AN EXEMPLAR-THEORETIC ACCOUNT OF WORD SENSES

R. E. Ramsey

PhD

2017

AN EXEMPLAR-THEORETIC ACCOUNT OF WORD SENSES

Rachel Elizabeth Ramsey

A thesis submitted in partial fulfilment of the requirements of the University of Northumbria at Newcastle for the degree of Doctor of Philosophy

Research undertaken in the Faculty of Arts, Design and Social Sciences

December 2017

Abstract

This thesis offers an exploratory study of certain aspects of the psychological status of the senses of four polysemous words, *over, under, above* and *below*, using three sets of sentence-sorting experiments. The thesis assumes that word senses are examples of linguistic categories; therefore, it is further assumed that categorisation tasks will offer insights into their nature as categories, and that effects predicted by existing models of categorisation can be tested. The first set of experiments questions the representativity of linguists' intuitions about the senses of these words. The results indicate that expert and naïve speakers' intuitions do not reliably coincide. This is consistent with existing research into the representativity of expert intuitions in syntax (e.g., Schütze, 1996). The possibility that this is due to individual differences is the subject of the second experiment. The data gathered suggest that there may be individual differences in word senses, consistent with observations of individual differences in other areas of language (e.g., Bates et al., 1995; Street and Dąbrowska, 2014). It is noted that lack of consensus may be a product of the task design, and that the scale of the task may have caused fatigue, forgetting, or semantic satiation. This was remedied in the final set of experiments. Further evidence of individual variation in word senses was found. In addition, the versatility of the methodology was exploited to test whether word senses are stored in memory, and in a manner compatible with the Generalised Context Model of Classification, or exemplar model (Nosofsky, 1986). Participants sorted the same stimuli twice, divided by a period of two months. In general, participants reached better consensus with themselves than with others. This indicates that word senses may have some form of mental representation. Effects of selective attention, a central prediction of the exemplar model, were observed in sorting behaviour. Four original contributions to knowledge are made: (1) there appear to be individual differences in word senses; (2) expert intuitions about what the senses of a given polysemous word are do not correspond to those of other speakers; (3) word senses do appear to have some form of mental representation, but not in the fixed form previously suggested (e.g., Tyler and Evans, 2001); and (4) selective attention effects are observed in this example of linguistic categorisation. The findings indicate that the exemplar model can account for the representation of word senses. This allows the conclusion that we may understand word senses as potential categories of exemplars.

Contents

List of tables

List of figures

List of appendices

Acknowledgments

I am indebted to my supervisors, Ewa Dąbrowska and Amanda Patten, for generously sharing their knowledge, advice and wisdom with me over the last three years. Without their consistent and solid guidance, and without their timely insights, this piece of research would not be what it is today. They have gone beyond their supervisory duties, too, and offered me a range of opportunities that have ensured my development as an academic. Helping to organise the 13th International Cognitive Linguistics Conference, and being entrusted as a module tutor, are two particular highlights. I am very grateful for the support and kindness they have shown throughout my PhD. I also wish to thank my former supervisors, Christopher Hart, and Christina Schelletter, for their guidance at the early stages of my research. I am also grateful to Christopher for introducing me to cognitive linguistics during my undergraduate degree.

My family and friends have shown me the love, encouragement and support I needed to begin, and reach the end of my studies. Loved ones in Hertfordshire, Newcastle and Sardinia have shown continued interest in my research, and they should know how motivating this is. They have also provided the distractions that have kept me sane during the course of my research. I am particularly grateful to my mum, Sandra, and sisters, Louise and Samantha, for encouraging me to move to Newcastle to take on this piece of work, and for their continued support prior to, and during my studies. In the second half of my studies, Chris gave me the encouragement and sustenance I needed to stay the course.

This research would not have been possible without the involvement of the more than five hundred members of the public who gave their time to participate in my experiments; without a generous studentship from the Faculty of Arts, Design and Social Sciences at Northumbria University; and without a software license generously provided by OptimalWorkshop.

This thesis is dedicated to my mum, Sandra, and grandad, Paddy, who inspired my love of learning, and to Freya, whom I hope I can similarly inspire.

Declaration

I declare that the work contained in this thesis has not been submitted for any other award and that it is all my own work. I also confirm that this work fully acknowledges opinions, ideas and contributions from the work of others.

Ethical clearance for the research presented in this thesis has been granted. Approval was sought from and granted by the Faculty of Arts, Design and Social Sciences Ethics Committee on 16 January 2014, and 11 March 2015.

I declare that the Word Count of this Thesis is 77,778 words.

Name: Rachel Elizabeth Ramsey

Signature: Rachel Ramsey

Date: 13 December 2017

Chapter 1 Introduction

Part 1

1.1 Polysemy and word senses

Polysemy, the phenomenon in which a word – indeed, most words (Murphy 2004; Clark 1983) – has a number of distinct, but arguably related senses, does not appear to present any particular problems to successful communication. What is troublesome about polysemy, instead, is how we go about *explaining* it – for example, the way in which we access a particular sense, the degree of interrelatedness of senses, and the representational status of senses remain open questions. For this reason, polysemy is the focus of intense and extended study in linguistics generally, and in cognitive linguistics in particular. This thesis aims to create new knowledge about polysemy, operating within the cognitive linguistic theoretical framework. I focus in particular on word senses, and certain aspects of their psychological status.

The overarching purpose of the thesis is to answer some questions that arise from consideration of some examples of polysemous words. In the tradition of cognitive linguistic research, this thesis examines the word *over*, along with *under, above,* and *below*. Let us consider some examples of *below*:

1. The sales value is well *below* target
2. Don't paint *below* the windowsill

I would predict that you, my reader, would agree that *below* exemplifies a different sense in each sentence. In my view, *below* in example 1 has a metaphorical sense and describes a position on a numerical scale. In contrast, example 2 locates a point in space. But what of example 3?

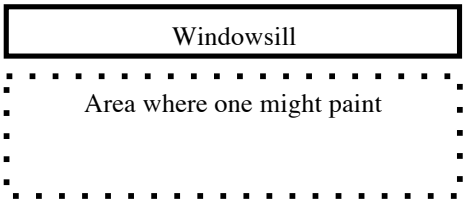


Figure 1 Schematic representation of *Don't paint below the windowsill*

3. Your mates are down *below*, watching you.

The difference between metaphorical and spatial senses of *below* is fairly obvious. Where the line between senses *within* these two broad categories should be drawn, however, is less clear. In this case, I would argue that examples 2 and 3 represent different senses; example 2 describes a two-dimensional relationship between two objects, which is illustrated in Figure 1, while example 3 represents a three-dimensional configuration in which the relative locations of the *mates* and *you* is on a diagonal, rather than vertical, axis, as illustrated in Figure 2.

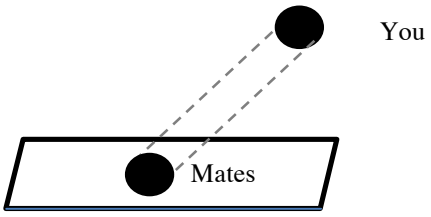


Figure 2 Schematic representation of *Your mates are down below, watching you*

This chapter introduces the overarching aims of the thesis, the approach taken to achieving those aims, fleshes out the concepts of polysemy and word senses, and introduces two theoretical accounts of categorisation. The chapter is divided into two parts, as follows. In part 1, following this brief introduction to the topic, I set out the aims of the thesis, and the research questions that are used to achieve those aims. I then describe the approach I take to answering those questions, and acknowledge the assumptions I make in taking this approach. This part of the chapter closes with a summary of the purpose of each of the following chapters, and how they interrelate. In the second part, I briefly describe the cognitive linguistic framework and the particular principle therein that guides this research. I then provide a brief evaluative summary of research in polysemy and word senses. This summary frames polysemy as a case of linguistic categorisation, and before this chapter closes, I introduce the two main theories of categorisation that have been invoked to account for linguistic categorisation: prototype theory, and exemplar theory.

1.2 Aims

In this thesis I aim to address four questions about word senses, which I answer with original data:

First, I ask whether the sense distinctions that I, as a scholar of the meanings of these four words, find meaningful coincide with those found meaningful by other native speakers of English. The answer to this question will create new knowledge about

the status of expert intuitions in the study of word meaning, and will contribute to a literature, principally concerned with intuitions about syntax (e.g., Dąbrowska, 2010; Schütze, 1996), about how representative expert intuitions about linguistic phenomena are.

Second, I ask whether native speakers of English agree with each other about whether a set of examples of a particular polysemous word use the same or different senses of that word. In this way, I aim to establish whether there are individual differences in word senses. The answer to this question will create new knowledge about individual differences in word senses, and will contribute to a large literature on individual differences in other aspects of language, such as grammatical attainment (e.g., Street and Dąbrowska, 2010), language acquisition (e.g., Bates, Dale, and Thal, 1995) and metaphor interpretation (Duffy 2015).

Third, I ask whether there is evidence that word senses are stored in memory. The outcome to this line of investigation will create new knowledge about the representation of word senses and will contribute to the unresolved debate over whether polysemous words are disambiguated by using context to "flesh out" a highly abstract meaning, à la the monosemy position (Ruhl 1989), or by accessing a sense stored in memory, as has been proposed by polysemy advocates such as Tyler and Evans (2001).

Following on from this third question, I ask whether the representation of word senses can be explained in terms of the (Generalised) Context Model of Classification (Medin and Schaffer, 1978; Nosofsky, 1986), which is frequently referred to by cognitive linguists as the exemplar model of categorisation. I assess this by testing a central prediction of the GCM, namely the effect of selective attention in making categorisation decisions[1]. In answering this question, I will contribute to existing exemplar-theoretic models of polysemy, building on work by scholars such as Gries and Divjak (2009). To date, linguistic applications of the model have typically shied away from addressing this aspect of the theory, or have

[1] Given the complex nature of selective attention, I offer a full explanation of the concept of selective attention in chapter 4.

implied that it may not be present in linguistic categories (Bybee, 2006, p. 716). While a limited number of scholars have predicted – and, where they have tested for them, indeed found – selective attention effects in language (Ellis, 2006; Francis and Nusbaum, 2002; Kalyan, 2012; Lively, Logan, and Pisoni, 1993), to my knowledge this effect has not been studied in polysemy.

If we return to the three examples of *below* presented earlier, I can offer a simplified summary of these four questions as follows:

1. Do you, my reader, and I agree about whether or not those examples exemplify different senses of *below*?
2. Do you and another reader agree with each other?
3. If asked to make the same judgment again in two months' time, would you reach the same conclusion?
4. Did my presentation of example 1 affect your judgment about whether examples 2 and 3 exemplify a different sense of *below*?

1.3 Approach and assumptions

Now that I have set out the aims of this research, I now outline my approach to answering these research questions, and specify the assumptions I make.

I assume that the senses of polysemous words are examples of linguistic categories. For that reason, I further assume that category members (i.e., examples of those senses) will behave in a manner consistent with how members of non-linguistic categories will behave. Principally, I assume that it is possible to organise stimuli into groups according to a given categorisation criterion; in this case, the meaning of a particular polysemous word, or target word, in context. I therefore assume that native speakers of English, acting as research participants, will be able to identify commonalities in meaning represented by uses of *below*, for example, and be able to organise into a single group all sentences in which the meaning of *below* is the same. Likewise, I assume that they will be able to identify differences in meaning, and that examples in which the meaning of *below* is different will be organised into separate groups.

Making that assumption, in this research I therefore adopt a categorisation task, operationalized as a sentence-sorting task. I use two variations of a sentence-sorting task. I use a closed-sort task to study whether the sense distinctions that I find meaningful are also found meaningful by other native speakers. I use open-sort tasks to study whether there are individual differences in word senses, whether they are stored in memory, and whether word sense disambiguation displays selective attention effects. In a closed-sort task, participants are presented with stimuli, in this case sentences, and predetermined groups, in this case what I judge to be different senses of the target word. They must categorise all sentences into one or more sense groups. In an open-sort task, participants are presented with sentences as stimuli, but do not receive predetermined categories. Instead, they must create their own categories to capture sentences in which the meaning of the target word is the same.

Sentence-sorting approaches have a firm foundation in the study of human word sense disambiguation and cognitive linguistic studies of polysemy (e.g., Cuyckens, Sandra, and Rice, 1997; Sandra and Rice, 1995). In the present research, I exploit the versatility of sentence-sorting tasks and combine them with uncommon statistical analyses to create a novel methodology for studying a number of aspects of the psychological status of word senses. In this way, I aim to not only contribute to knowledge about the status of word senses, but offer a novel tool for understanding certain aspects of their nature.

1.4 Summary of chapters

This thesis describes an exploratory and empirical study of particular aspects of word senses, in which each experiment builds upon the findings of its predecessor. It is comprised of five chapters, the purpose of which is summarised below.

Chapter 1. Introduction

I use this chapter to introduce the topic of the thesis, specify the aims, approach and assumptions. It then provides an introduction to the cognitive linguistic framework and its principles that guide this research, before fleshing out the concepts of polysemy and word senses in more detail. The chapter then presents the theoretical background underpinning the thesis by introducing and comparing two major theories of categorisation.

Chapter 2. Experiment 1: A closed sentence-sorting study to test the representativity of linguists' intuitions about word senses

I describe in this chapter the first of three sets of experiments concerned with particular aspects of word senses. Specifically, I report a study challenging the representativity of expert intuitions about word senses, by testing whether my intuitions about whether a particular example of a polysemous word is an example of a particular sense of that word are shared with other native speakers of English. The chapter consists of two parts: in the first part I provide a critical review of existing research on the role and status of expert intuitions in the study of linguistic phenomena in generally, and in polysemy in particular. In the second part I describe a closed sentence-sorting task completed by 298 native speakers of English. Before closing, I identify a line of further inquiry that is needed to substantiate my interpretation of the data.

Chapter 3. Experiment 2: An open sentence-sorting task to test for individual differences in word senses

In this chapter I report the second of three experiments. It is the purpose of this experiment to follow up on a finding made in the first experiment. Specifically, it tests whether the outcome of the first experiment can be explained on the grounds that individuals have different senses of the polysemous words *over*, *under*, *above* and *below*. Again, this chapter comprises two parts. The first part consists of a critical review of existing research on individual differences in language generally, and in word senses in particular. The second part describes a large-scale open sentence-sorting task completed by 44 native speakers of English. I close the chapter with some comments concerning impact the design of the experiment may have had on the way participants completed them, and identify how the experiment may be improved.

Chapter 4. Experiment 3: Testing an exemplar model of word senses

In this chapter I report the final experiment, which builds upon the approach taken in Experiment 2. In light of my comments concerning the impact of the experiment design on participants' responses, and with the aim of gathering more reliable data, I report an iteration of the open sentence-sorting task used in Chapter 3. In this study,

the task is significantly smaller in scale. The experiments reported follow on from the findings of Chapter 3 in that they too investigate the possibility of individual differences in word senses. The versatility of the methodology is exploited further in this chapter, and the experiments are also used to shed light on the question of whether word senses are stored in memory or created ad hoc, and, if they are stored, whether their storage is compatible with a central prediction of the (Generalised) Context Model (Nosofsky 1986; Medin & Schaffer 1978), also known as the exemplar model. Specifically, I test whether participants' responses in these linguistic categorisation experiments are subject to selective attention effects. The chapter again comprises two parts. The first part provides a critical review of literature on categorisation in general, and the application of psychological models of categorisation to account for linguistic categories. The second part comprises a report on a series of open sentence-sorting tasks completed by 205 native English speakers.

Chapter 5. General discussion and conclusions

This final chapter serves to bring the thesis to a close. I summarise the principal findings of the three sets of experiments, and discuss their theoretical and methodological implications. I confirm that the aims of the thesis have been met, and present the four primary original contributions to knowledge. I acknowledge the limitations of the research, and set out some suggestions for future investigation.

Part 2

This part of the chapter situates the research in a broader theoretical context, identifying both the theoretical framework in which this research operates, and the theoretical explanations given to account for polysemy to date, and presents a concise overview of polysemy and word senses.

1.5 Cognitive linguistics: the generalisation and cognitive commitments

This research operates in the theoretical framework of cognitive linguistics. This framework is relatively young, yet aims to provide cognitively realistic explanations of all linguistic phenomena. Contrary to traditional linguistic theory, cognitive linguistics argues that language is not a modular function distinct from other cognitive domains. Instead, cognitive linguistics claims that language is just one of a number of interrelated cognitive functions. For that reason, theorists in the field argue that language is processed, accessed, stored, and so on, in the same way as other cognitive phenomena. This entails that cognitive linguists aim to explain linguistic phenomena in a manner consistent with what is already known about the mind more generally. This entailment was described by Lakoff as the cognitive commitment, which requires that we should "make [our] account of human language accord with what is generally known about the mind and the brain, and from other disciplines as well as our own" (1990, p. 40). This thesis aims to adhere to this commitment. Lakoff advocated an additional commitment, the generalisation commitment, which recommends that cognitive linguists should aim to characterize the "general principles" which underlie all human languages (p. 40).

Of these two commitments, it is the cognitive commitment which is of particular relevance to this project. Indeed, this commitment underpins most research on polysemy and word senses in cognitive linguistics, for this work often considers polysemous words to be examples of linguistic categories (e.g., Brugman and Lakoff, 2006 [1988]; Brugman, 1981; Klein and Murphy, 2002; Rice, 1993; Taylor, 2003; Tyler and Evans, 2001). For this reason, accounts of the representation of polysemous words and their senses draw heavily on research from cognitive

psychological models of categorisation. Specifically, cognitive linguists have traditionally favoured prototype-based categories developed by Rosch and her colleagues (Rosch and Mervis, 1975; Rosch, 1973), and have drawn on this model when positing prototypical senses and the organisation of word senses. That said, exemplar-theoretic models of word meaning are growing in currency, in keeping with work in other areas of cognitive linguistics that advocates for exemplar theory-compatible representations. These models will be discussed in more detail later in this thesis. The cognitive commitment should also entail that we study language using tools and methodologies borrowed from cognitive science. As I will discuss later in this thesis, this is not always the case, and expert intuitions remain key features of the cognitive linguist's toolkit.

1.6 Polysemy

Polysemy is a major concern within cognitive linguistics. Cuyckens and Zawada (2001, p. xv) note that polysemy is described as being "rampant" in the field, with researchers claiming that infinitely fine-grained senses are related to a central sense. A polysemous word is one that has a number of different, but related, senses. Polysemy and homonymy (the phenomenon whereby words which share the same form have unrelated meanings) therefore contrast, for example:

4. a. The house rests *on* the foundation.
   b. He lives *on* a pension.
5. a. They put the dog in the *pound*.
   b. John has lost one *pound* this week.

In example 4, taken from Beitel, Gibbs, and Sanders (1997), the two uses of *on* share some common characteristics. While example 4a relates to a specifically spatial configuration, one in which the house (figure) and foundation (ground) have a vertical relationship with contact between them, with the ground providing physical and structural support to the figure, the use of *on* in example 4b is more abstract in nature. According to Beitel et al., though, the function of the foundation in 4a corresponds with the function of the pension in example 4b; while the foundations provide physical support, the pension provides financial support. The use of *pound*

in examples 5a and 5b, in contrast, share no commonality; the nature of an official enclosure for dogs has no relationship with the nature of a unit of mass.

A theory of word meaning that argues in favour of polysemy therefore rejects the monosemy hypothesis developed by Ruhl (1989). Ruhl claims that "a word has a single general meaning," (p. 234); moreover, he argues that "a word should always be assumed to contribute as little meaning as possible to its context" (p. 8). Meaning, therefore, results from the interaction between collections of these underspecified units in sentential context, and between these units and extra-linguistic information. The monosemy theory therefore rejects any notion of distinct senses represented in memory.

1.7 Senses of polysemous words

In polysemy literature, *senses* of polysemous words contrast with *meanings* of homonyms; whereas *meanings* of homonyms are distinct from one another, *senses* of polysemous words are distinct, but are argued to be related[2] (e.g., Tyler & Evans, 2001). There is, though, more to be said about the nature of this phenomenon. According to Hanks (2000), senses (and indeed meanings of homonyms) are nothing more than meaning potentials: a bundle of components that are activated in isolation or combination according to the surrounding context. A dictionary, on this basis, is therefore an inventory of meaning potentials associated with each form. Though this might sound a little like the monosemy hypothesis described by Ruhl (1989), it is important to note that where proponents of the monosemy hypothesis claim that most words have a single, general and very abstract sense which is substantiated by context, Hanks does not suggest here that words have a core meaning but instead have a set of components that are available for contextual activation. In these terms, then, a sense of a polysemous word can be understood as a particular component or combination of components. In dictionaries, the example given under a sense is therefore an example of the context that might activate the component(s) associated with that sense. This account is consistent with other descriptions of polysemous word senses; where Hanks refers to the component(s) underlying each sense,

[2] Though Klein and Murphy (2002) find that senses of some polysemous words may share very little common meaning.

Brugman and Lakoff (2006 [1988]) and Tyler and Evans (2001) refer to the meaning components – such as verticality and contact – associated with each sense.

Kilgarriff argues that "an individual's history of hearing a word dictates his or her understanding of that word" (2007, p. 37–8). Such an argument allows the conclusion that if two individuals' histories of hearing a word are different, their understanding – and therefore their senses – of that word are also likely to be different. This prediction is consistent with the cognitive linguistic conception of language as a usage-based phenomenon (Barlow and Kemmer, 2000). Such a conclusion explains why subjects completing word sense disambiguation (WSD) tasks often fail to reach complete agreement (e.g., Passonneau, Salleb-Aoussi, Bhardwaj, and Ide, 2010). The possibility that individuals may disagree over what sense of a polysemous word a given example exemplifies has not been explored in detail by cognitive linguists, and seminal work on the topic, such as that by Brugman and Lakoff (2006 [1988]) and Tyler and Evans (2001), implicitly assumes no individual variation in word senses.

As noted above, the representational status of word senses is uncertain. On the one hand, scholars such as Ruhl would contend that, since words are disambiguated using sentential and environmental context, the meaning of a given word in context is not stored in memory. On the other, polysemy scholars such as Tyler and Evans (e.g., 2001) argue that at least some senses are stored in memory.

1.8 Polysemy as a case of linguistic categorisation

Within the cognitive linguistic framework, polysemous words are typically held up as examples of linguistic categories (e.g., Taylor, 2003). This account works particularly well for scholars advocating a system of senses organised around a prototypical sense, such as that which has been proposed by Brugman and Lakoff (2006 [1988]) and Tyler and Evans (2001). For example, Tyler and Evans state that scholars "have argued that lexical items constitute natural categories of related senses organised with respect to a primary sense and thus form semantic or polysemy networks." (2001, p. 726). Canonical analyses of polysemous words, such as those offered by the authors just mentioned, draw on the prototype model of categorisation (e.g., Rosch and Mervis, 1975; Rosch, 1978) when describing the

organisation of the senses of polysemous words. Their accounts are not true copies of the prototype model; the prototype model states that the category can be represented by a prototype, which itself is an abstraction of the characteristics of all category members. In contrast, the models proposed by Brugman and Lakoff and Tyler and Evans assume an abstract but meaningful prototypical sense which, when combined with cognitive principles such as reconceptualization, derive a set of related senses. More recently, accounts of polysemy that are more aligned with the exemplar theory of categorisation have been put forth (Gries 2006; Gries 2015).

Whichever theoretical model is used to account for polysemy, it remains that polysemous words are generally understood in the cognitive linguistic framework as linguistic categories, and senses as members of those categories. It is unclear why the possibility that word senses may instead be categories, and example sentences members of those categories, has not been explored. If this were the case, in the context of this set of experiments we would expect to see categorisation effects taking place at the level of senses. Chapter 4 of this thesis sets out to establish whether this is indeed the case, and whether a particular categorisation effect, specifically that of selective attention, is observed in the sense categories that participants make.

Given the proposition made by cognitive linguists – and indeed this thesis – that polysemous words, or their senses, are examples of linguistic categories, in the following section I step back and introduce categorisation as a more broad concept and present two competing theories of categorisation, both of which have been invoked to account for polysemy.

1.9 Theories of categorisation

Human beings are immersed in an environment saturated by potential stimuli; we are capable of attending to inordinate volumes of tastes, motor experiences, sounds, visual information and smells. It is essential that we are able to make sense of these stimuli in order that we can make assumptions about them and, where appropriate, modulate our behaviour around them. For example, if we were to encounter the creature in Figure 3, our ability to accuractely recognise it as a DOG means that we

can confidently assume that they will require food and water, they may shed fur on furniture, and that we should think carefully before introducing it to a cat.

Figure 3 Labrador (Wikipedia, 2008)

Despite the scale of this sensory overload, it does not appear that humans struggle to manage it. It is proposed that humans make sense of their surroundings by means of categorisation. The traditional, Aristotelian account of categorisation proposed that categories can be defined in terms of an object having a set of individually necessary and jointly sufficient attributes. An object must feature all necessary and sufficient characteristics in order to achieve category membership; membership is therefore an all-or-nothing affair. A number of problems with the classical model have been raised. First, the empirical reality of typicality effects problematizes the proposition that all category members meet necessary and sufficient criteria. If that is so, then all members must be equal. It has been observed that category membership is subject to subjective gradation, with particular examples being judged as more typical than others (Rosch et al., 1976). Second, some categories are difficult to define. A famous example is that of the category GAME (Wittgenstein 1958). Football, chess, solitaire and hopscotch are all games, but which features unite and therefore define them? Third, there is the issue of variability in judgments of category membership, such as whether a tomato is a fruit or a vegetable.

Contemporary theories of categorisation aim to address the inadequacies of the classical model, and the rest of this chapter is dedicated to introducing two of them: prototype theory, and exemplar theory.

1.9.1 Prototype theory

Prototype theory, developed by Rosch and her colleagues (Rosch et al., 1976; Rosch & Mervis, 1975; Rosch, Simpson, & Miller, 1976; Rosch, 1973, 1978), offered a response to the above-mentioned problems with the classical account of

categorisation. Their account divided categorisation into two dimensions: the vertical dimension, representing category inclusivity, and the horizontal dimension, representing category distinctiveness. The two sections that follow outline these two dimensions.

1.9.1.1 The vertical dimension: inclusion

Rosch (1978) proposes that categories are organised in a vertical dimension of inclusivity. The category at the top of this structure – the superordinate category – is that which is the most inclusive, and which offers a generic term that can be applied to all categories that it encompasses. While it is the most inclusive level, in feature terms it shares little in common with the categories under it; as Rosch and Mervis (1975) note, the features that superordinate categories share with its basic and subordinate level categories are typically abstract. For example, across the entire category members of the superordinate category FURNITURE shares the attributes of being physical objects, and being objects that are found inside buildings, on the whole they will share little else. Beneath the superordinate level is the basic level; in contrast with the superordinate level, category members at the basic level have a high degree of featural overlap. For example, members of a basic level category such as TABLE will share many features, such as having legs, and a flat surface. Beneath the basic level exists the subordinate level, comprising more specified versions of basic level categories, such as COFFEE TABLE, and DINING TABLE. Rosch and Mervis propose that basic level categories have highest cue validity as a result of their simultaneous high degree of featural overlap within members of the category (absent in superordinate categories) and low degree of featural overlap across categories (present in subordinate level categories). Basic level categories therefore have maximal between-category heterogeneity and maximal within-category homogeneity.

While one might intuitively predict that the most generic and inclusive level – i.e., the superordinate level – is most basic, empirical evidence indicates that this is not the case. Evidence for the priority of the basic level has been drawn from the studies of first language acquisition, motor programs associated with objects at each level of abstraction, object recognition, object naming, and linguistic evolution (Rosch et al., 1976).

While it is proposed that taxonomic categorisation is universal, membership at each level is subject to some degree of individual variation as a consequence of varying subjective experiences and interactions with the world around us. This variation exists both within and between linguistic communities. One of the subjects involved in Rosch et al.'s (1976) investigation into basic level categories had particular technical expertise, experience and familiarity with aeroplanes. It was argued that it was this extra-developed understanding of the aeroplane category that produced results indicating that aeroplane was, for this subject, a superordinate category instead of a basic category; basic categories were instead specific instances of aeroplanes. Offering a cross-linguistic insight into taxonomic categories, Berlin, Breedlove and Raven (1974) reveal that for speakers of Tzeltal in the Tenejapa region of Mexico, basic level categories exist at the GENUS-level of a folk-classification taxonomy (genus examples include oak trees and maple trees), whereas basic-level categories in the minds of Britons are likely to be closer to the LIFE FORM-level (tree, for example), an interesting observation that becomes more exciting when one considers that the GENUS- and LIFE FORM-levels are, for Tzeltal speakers, separated by a further, INTERMEDIATE level (which includes leaf-bearing and needle-bearing trees).

Cognitive economy is, according to Rosch, maximised by this approach to category organisation. Priority of the basic-level category is maximally informative in a way that might be compared with the Goldilocks fairy tale. The superordinate level is highly abstracted, and provides very little information about the category. The subordinate level, in contrast, is over-specified and provides excessive information typically relevant to only a small subset of the category. At the basic level, the level of detail about the category is just right: enough information is provided to distinguish it from other basic level members of other categories, and enough information is provided to allow subordinate category members to be associated with it.

1.9.1.2 The horizontal dimension: differences and prototypes

As Rosch (1978) observes, the world as humans perceive it "is not an unstructured total set of equiprobable cooccuring attributes." (p. 4). Instead, there exists

correlational structure in object attributes; to use the frequently-used example, feathers co-occur with more frequency than they do fur, which allows us to judge that if a creature has fur, it is unlikely to have wings. Correlational structure therefore affords contrast that can be used to categorise a novel object; if a novel object has feathers and we do not have time to establish anything more than that, we can categorise that object with a reasonable degree of confidence as a BIRD. This dimension of difference is most effective where it intersects the basic level of the vertical dimension. At the basic level, the extent of informativeness maximises differences between categories; for example, the characteristics of the basic level category CUP differentiate it from another basic level category such as CAR. At the subordinate level, this dimension still allows differentiation, but to a lesser degree; a TEA-CUP is less different to a MUG than a CUP is to a CAR.

This dimension of difference between categories licenses category prototypes, which are typically – though as I will come to, not universally – understood as an abstraction of the features of category members; in this way, it captures a 'summary representation' of the category.

1.9.1.3 The nature and purpose of prototypes

Rosch proposes that categorisation is achieved by means of comparing a novel stimulus to the prototype of a candidate category. Just what a prototype is, however, is the topic of debate between authors, and even within single publications. For example, Rosch and Mervis (1975, p. 575) indicate that a category prototype is a representation consisting of average features of category members ("we view semantic categories as networks of *overlapping attributes*" - my emphasis). It seems likely that it is this statement that Murphy (2004, p. 41-2) refers to when he says that Rosch and Mervis "explicitly deny" the interpretation that "every category is represented by a single prototype or best example". However, since Rosch and Mervis state *on the same page* that prototypes can be more broadly defined as "those category *members* to which subjects compare items when judging category membership," (my emphasis) it seems unsurprising that some readers have understood prototypes as being a particular member of the category. Indeed, this is the view taken by Croft and Cruse (2004, p. 77) when they describe prototypes or prototypical members as "best examples of categories". They do, however, later (p.

81-2) state that two versions of prototypes exist, one which posits an average of category features, and one which posits a specific member as the prototype; they note that linguists have been guilty of failing to distinguish between the two versions. Taylor (2003) makes similar comments, this time arguing that as well as the average features and specific item (in his words, "exemplar") accounts of prototypes, a third exists, which posits that certain types of category members can be prototypes. He later states that he adopts the individual exemplar view of prototypes in the remainder of his monograph on linguistic categorisation (p. 69). In keeping with the interpretation most prevalent in the literature, this thesis operates on the understanding that prototype-based models assume that the category is represented by a summary representation of all members of that category, and that *that* is the category (Murphy 2004, p. 42).

As mentioned, categorisation is achieved by comparison of a novel stimulus to a category prototype. However, not all characteristics or features of the prototype have equal status. Instead, some features are weighted, with this weighting capturing the frequency with which a given feature is associated with members of the category, and the distribution of those features across contrasting categories (Rosch and Mervis, 1975). Features which are most frequent across members of the category are most highly weighted, and those which are least frequent have low weighting. Likewise, features that are more frequent in contrasting categories will have lower weighting than those that are less frequent. Assignation of a novel stimulus to a given category is most likely when that stimulus possesses the category prototype's high-weighted features (Murphy 2004).

1.9.1.4 Family resemblances

As we have seen, differences in the distribution of features distinguish the categorisation levels on the vertical dimension. Likewise, differences in the distribution of features are what distinguish prototypes on the horizontal dimension. In this way, differences in feature distribution determine the structures underpinning categories at a general level. However, Rosch and Mervis (1975) also propose that differences in the distribution of features determine the structure of categories at a more specific level. Specifically, they propose that such differences explain why some members of a category are judged to be more typical of the category than

others. The term *family resemblance* is used to describe degrees of featural overlap in category members; category members that share more common features will have closer family resemblance than those that share fewer common features. Moreover, those that have stronger family resemblance to all other members of the category will be considered more prototypical than others. Degrees of prototypicality are also affected by the category member's degree of family resemblance with members of other categories: the lower the featural overlap and degree of family resemblance with members of contrasting categories, the more prototypical of the category the member will be. Rosch and Mervis propose that these varying degrees of family resemblance, resulting in varying degrees of prototypicality, can be understood spatially. In a spatial representation, centrality represents prototypicality, and increased family resemblance with other category members results in increased proximity to the category centre.

1.9.1.5 Prototype categories: summary

Rosch and her colleagues' work represented a step change from classical accounts of categorisation, providing a framework that could account for gradation in category membership, that could explain why both solitaire and hopscotch are games, despite their limited featural overlap, and that could account for variation in judgments of category membership. At the centre of this account is the prototype, which acts as the category representation, and the point of comparison called for when categorising a novel stimulus. Typicality effects, which were a snagging point in the classical account, are accounted for by prototype theory by the proposition that category members can be measured for their family resemblance to other category members. Those that share most features with other category members will be considered more typical – and in spatial terms, more central – than those that share few features with other members. These members will, in spatial terms, be more peripheral members of the category.

1.9.2 Exemplar theory

While prototype theory proposes that a category is represented by a prototype capturing a summary representation of all members of the category, exemplar theorists propose that categories comprise tokens of previously encountered exemplars. In this way, prototype and exemplar theories occupy opposing ends of a continuum of category abstractness; while prototype theory endorses a highly

abstract category representation, exemplar theory claims that the category has no abstraction. Categories comprise exemplars that are similar to each other; in exemplar theory, similarity is modelled in terms of spatial proximity. Exemplars that are highly similar to each other are positioned closely together, while exemplars that are highly distinct from each other are positioned far apart. Categories therefore comprise exemplars that are in close proximity. In this theory, exemplars are understood to occupy a multidimensional psychological space (MDS), in which dimensions correspond to features. The exact location of an exemplar in the MDS is determined by the value of the exemplar on each feature dimension.

An early and influential account of exemplar theory was proposed by Medin and Schaffer (1978) in their Context Theory of Classification Learning, and was later generalised by Nosofsky (1986). This model, according to Murphy (2004, p. 89), has served as the basis of most popular models of exemplar-based categorisation. In their paper, Medin and Schaffer specify a number of assumptions about the model, including the assumption that a novel item *i* is categorised to a category *j* on the basis of similarity between *i* and exemplars of *j*, and between *i* and all exemplars. As noted above, exemplars are positioned in a multidimensional space, with their precise location corresponding to the value of each variable dimension. Importantly, the dimensions of this space – and therefore the (potential) categories within it – are flexible. In an exemplar-based account of categorisation, Medin and Schaffer, and Nosofsky propose that in a categorisation task, the categoriser *selectively attends* to a particular (set of) characteristic(s). Given that this concept of selective attention is central to this study, I provide in section 1.9.2.1 an example of two categorisation scenarios, and afterwards explain how the exemplar model explains the different approaches to categorisation taken in each. Following this, I move to discuss a second, more implicit prediction of exemplar theory: individual differences.

1.9.2.1 Selective attention

Imagine that an individual is sorting clean cutlery with different coloured handles into a drawer. In this case, accurate categorisation – i.e., assignation of an exemplar cutlery to the right area in the cutlery drawer – will depend on the individual correctly distinguishing the items based on particular characteristics. Of primary importance in a typical cutlery sorting activity is likely to be the shape of each

exemplar, with size (to distinguish between different types of spoons) and sharpness (to distinguish between steak knives and table knives, for example) being secondary concerns. Accordingly, the individual will selectively attend to these characteristics, and will overlook other characteristics, such as handle colour. Let us say that this individual later buys a replacement set of cutlery. Imagine a scenario in which her old cutlery is on the draining board, having been washed previously, and she is washing the new set before it is used for the first time. She does not intend to continue to use the old set on a regular basis, and she wants to keep it separate from the new set. The old set will be put at the back of the drawer. She then needs to implement an additional, primary classification criterion: old versus new. In this scenario, at the first stage, only this criterion is of interest, and therefore – at this stage – she no longer categorises the cutlery into different types. Those differences still exist, but are no longer relevant in the task of dividing the old and new cutlery.

An exemplar-theoretic account of this person's categorisation decisions posits that dimensions in a multidimensional psychological space (Medin and Schaffer, 1978; Nosofsky, 1986) represent the cutlery's characteristics, such as their shape, sharpness, handle colour, etc. Crucially, it is claimed that these dimensions shrink and stretch according to the goal of the categorisation task (Nosofsky, 1986, p.41), therefore changing the relative position of an exemplar in the space. In other words, categorisation under the exemplar model is inherently task-based, and categories are dynamic, unfixed entities. Specifically, the dimension of interest, i.e., that which is of most importance in the categorisation task, stretches, and irrelevant dimensions shrink. As noted above, similarity of exemplars is modelled in terms of spatial proximity, meaning that when a dimension stretches, the similarity of exemplars along that dimension effectively decreases; this has the effect of making distinctions between the items along that dimension being more apparent, which facilitates easier categorisation. When a dimension shrinks, similarity of exemplars along that dimension effectively increases, making the status of that feature in each of the exemplars less apparent. In the first cutlery-sorting scenario, the individual selectively attends to shape, size and sharpness. She does not attend to the colour of the handle, making this an irrelevant dimension. The dimensions of the relevant characteristics, namely shape, size and sharpness, will expand to reveal finer distinctions between the exemplars along these dimensions. The stretching of the

size dimension, for example, will therefore accentuate differences in spoon size. The dimension associated with the irrelevant characteristic, the colour of the handle, shrinks, making the different colours effectively less distinct. In the second scenario, the distinctions amongst the old cutlery still exist – there remain observable differences in shape, size and sharpness – but at this time, for categorisation to be accurate the individual need only attend to the age of the exemplars. Accordingly, this dimension of age expands revealing distinctions between the old and new cutlery, and the dimensions associated with shape, size and sharpness shrink. It is this stretching of relevant dimensions, which highlights differences, and shrinking of irrelevant dimensions, which obscures differences, that optimises categorisation accuracy.

1.9.2.2 Selective attention and contextual modulation

While the idea that category membership is dynamic is central to exemplar theory, it does not form part of the primary prototype theory literature. However, contextual modulation, in which category membership varies according to the context of the categorisation scenario, has been incorporated into later literature grounded in prototype theory. For example, Labov (1978) reports a series of experiments in which a single vessel could be alternatively called a *cup* or *bowl* depending on whether or not it contained mashed potato. Figure 5 shows the frequency with which the cups in Figure 4 are called *cup* or *bowl* in a neutral context (when empty), and in the food context (when filled with mashed potato). Figure 5 shows that vessel 4, for example, was always referred to as a *cup* regardless of context. In contrast, while in the neutral context vessel 2 was referred to as a *cup* around 85% of the time, and as a *bowl* around 15% of the time, when shown in the food context, the same vessel was referred to as a *cup* around 30% of the time, but as a *bowl* around 65% of the time. These findings suggest that the context in which a given vessel is seen plays some role in determining what it is labelled.

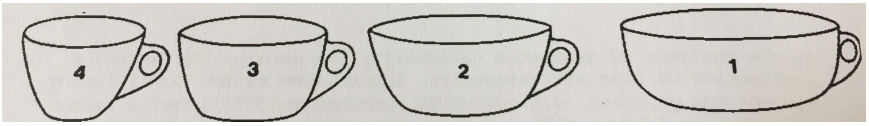


Figure 4 Vessels used in *cup* / *bowl* experiments (Labov, 1978, p. 222)

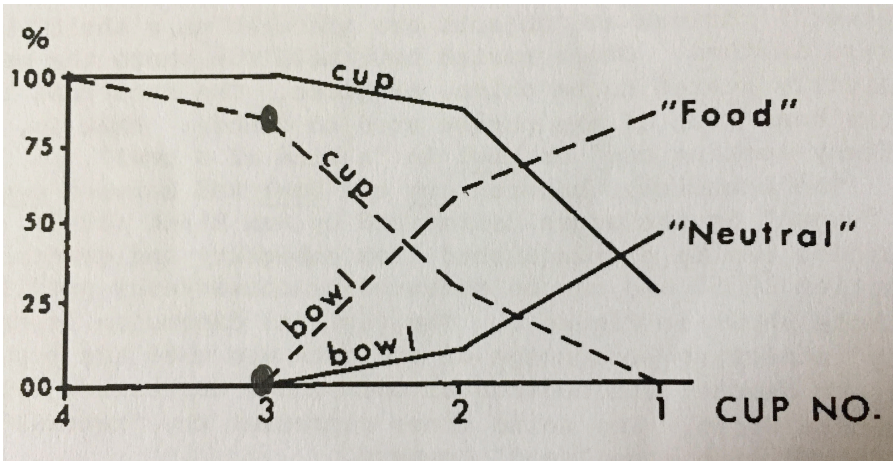


Figure 5 Consistency profiles for *cup* and *bowl* in neutral and food contexts (Labov, 1978, p.224)

In alignment with prototype theory, Labov proposes that the features of a stimulus – in this case, the size of the vessel, presence or absence of a handle, what it contains, and so on – are weighted. Where Labov progresses from prototype theory, however, is in his articulation of the interdependence of the features; the contribution a particular feature makes to the categorisation decision can vary according to the nature of other features. In Labov's example, regardless of the context his experiment participants encountered vessel 2, it had identical physical characteristics, each with the same weighting. However, when placed in the food context, the presence of mashed potato in the vessel adjusted the weighting of the features that would, in the neutral context, result in it more reliably being called a *cup*. The result is that more than half of the participants called the vessel a *bowl*.

For research that is grounded in prototype theory, it is surprising that Labov proposes that the weighting of a given feature can change. As noted above, prototype theory proposes that feature weighting is a function of the frequency with which a given feature occurs in members of the category, and of the distribution of that features across members of other categories. In this way, feature weighting can change, but only as a result of exposure to more members of the category, and to members of other categories. In this respect, Labov's proposal that weighting is dynamic is perhaps more closely aligned with exemplar theory, of which feature weighting adjustment forms an integral part.

1.9.2.3 Individual differences

While selective attention is a central component and testable prediction of the exemplar model, individual differences in categorisation is a more implicit prediction. It is implicit in that it is a logical outcome of selective attention, and the proposal that categories comprise memory traces of tokens of previously-encountered exemplars.

Selective attention entails that categories are created in response to a task demand. In a simple example, if one was required to categorise a set of differently-coloured blocks of a range of sizes by colour, one could attend to the colour of the blocks at the expense of differences in shape. Individual differences may be observed in the exact outcomes; for example, one participant may sort a turquoise block with blue blocks, while another may sort them with green blocks, while a third may set them apart from both green and blue blocks. In a more complex example, related to the present research, when asked to decide whether two uses of *over* to describe spatial scenes are the same or different, one could attend to a larger number of dimensions. For example, in the sentences *They keep slinging their towels over the bedroom door* and *I go over the handlebars*, one may attend to the arc-like trajectory captured in both sentences. An individual only concerned with attending to the arc-like trajectory might be satisfied that both sentences exemplify the same sense of *over*. A different individual, concerned with the ultimate position of the figure following completion of that arc-like trajectory, may consider them exemplars of different senses; one which describes the arc-like trajectory of a figure relative to a ground, and one which captures an arc-like *shape* of a figure at the end of the motion. These two examples, of varying complexity, serve to demonstrate that categorisation decisions (including linguistic categorisation decisions), as they are made on the basis of a categorisation criterion which may vary across individuals, are subject to individual differences.

The notion that categories comprise memory traces of previously-encountered exemplars also motivates the prediction that categories are subject to individual differences. Take jackfruit, for example. A British individual who has not encountered a jackfruit before, if required to establish what it is likely to be and make predictions about it (e.g., what kind of meal it would be eaten as; what would it would taste like) might compare it with other foods that feature the word *fruit*, or

indeed other members of a general FRUIT category. On that basis, they might assign it to the FRUIT category, predict that it will be sweet, and could be eaten as dessert. However, another British individual, this time one who follows a vegan diet, may have been encountered and eaten a jackfruit and assign it to a MEAT REPLACEMENT PRODUCT category, and know that it could be eaten as part of a savoury dish. A third individual, this time one from a location in which jackfruits are eaten regularly, such as Indonesia, may assign it to the same category as the first British individual (FRUIT), because they have eaten it in a way that other fruits are consumed. This example demonstrates that categorisation decisions, as well as being determined by individual decisions, are also influenced by experience with the category. In this way, exemplar theory implicitly predicts that categorisation decisions can differ systematically between individuals on a large scale.

1.9.3 Multiple categorisation processes

While research on categorisation typically contrasts the explanatory power of one model over another, a growing body of work exists which investigates how multiple models can be combined to explain observations about categorisation. For example. Smith and Sloman (1994) found that subjects categorised common objects based on rules *and* similarities. These findings were supported by later work by Sloman and Rips (1998). Elsewhere, it has been argued that a single model can accommodate the characteristics of the prototype and exemplar models, which respectively propose total abstraction and no abstraction. Vanpaemel and Storms (2008) propose that the Varying Abstraction Model can be used to accommodate not only exemplars and prototypes, but also intermediary representations, consisting of category members clustering to form subprototypes. According to this account, an intermediate representation has goldilocks characteristics: it is neither too sparse and lacking in useful detail, but nor is it too detailed and cognitively uneconomical. Only in extreme and necessary cases are full abstractions or exemplars represented in memory. This account can therefore be characterised as an abstraction continuum: at some point in category development, total abstraction is necessary, at others no abstraction is needed, and at others still only partial abstraction is needed. Baetu and Shultz's (2010) investigation of the processes underlying concept learning, while not acknowledging an intermediate representation, offers further insights on this theory. Their results suggest that, in cases where the concepts being learned do not possess

defining features, abstraction is initially used to form a concept prototype, against which novel items are compared. Over time, this tendency changes, and the model instead comes to rely upon comparison between old trained items and the novel item. The authors speculate that this can be explained by virtue of the fact that as time passes and experience with category members increases, trained examples are better remembered, and there is a consequential reduction on the reliance on comparison with a prototype. In effect, then, the model must reach a tipping point; before this, it abstracts information from novel items to form a prototype. At some point the model is sufficiently familiar with trained items that it becomes more efficient to compare novel items with items already experienced. The authors do not speculate about when that tipping point might occur, but nevertheless present an appealing proposal of how different levels of abstraction are used at different points of concept learning. Divjak and Arppe (2013) present further evidence supporting the integration of exemplar- and prototype-based models of categorisation. Using Russian and Finnish corpus data, they propose that repeated exposure to exemplary exemplars (that is, exemplars with strong association with a particular property in the mental lexicon) results in the abstraction of a prototype core. As such, the authors argue for a varied abstraction model, which does not prioritise one degree of abstraction (for example total, in the case of a prototype) over another.

1.10 Categorisation and polysemy: some areas for investigation

Sections 1.8 and 1.9 have served to set out the assumption that polysemy is an example of linguistic categorisation, and to describe the variability of the categorisation theory landscape. This landscape features two poles: to one side exists a theory that proposes highly abstract categories comprising a summary representation of its members; to the other, categories are proposed to comprise tokens of previously encountered exemplars. Between these two poles is an intermediate account that allows for abstraction *and* individual exemplars.

As already noted in section 1.8, canonical accounts of polysemy have attempted to explain the phenomenon in prototype-theoretic terms; while there is not an exact match between the prototype theory proposed to account for semantic categories and the version used to account for polysemy, there are clear overlaps, for example in the

invocation of prototypes and the degree of abstraction in the category. While more recent work on polysemy has explored the potential that exemplar theory might have in accounting for polysemy, it forms a small literature, and certain predictions of the model have not been tested.

While prototype and exemplar theories have overlaps – both acknowledge the reality of varying degrees of typicality within categories, for example – they diverge in a way that allows one to test whether the predictions of either model are observed in this case of linguistic categorisation. Of particular interest are the predictions made by exemplar theory; the implicit prediction that polysemy, like other types of categorisation, will be subject to individual differences; and the explicit prediction that categorisation decisions are task-specific, and that categories emerge in response to a categorisation criterion. While prototype-theoretic research by Labov (1978) has explored contextual modulation in categorisation decisions, his explanation of how such modulation occurs – that the weighting of a given attribute varies according to context – runs contrary to prototype theory proper, which does not propose varying feature weighting. This aspect of his work is more closely aligned with exemplar theory, of which the idea that feature weighting varies by context is a central component. While prototype-theoretic research has briefly explored variation in categorisation, it has explained such variation on the basis of cultural differences (e.g., differences in cuisines), and differences in degree of knowledge (e.g., differences in extent of familiarity with category members, resulting in differences in what type of members exist at the superordinate, basic and subordinate category levels). Such an explanation would have limited power should individual differences be the norm in the categorisation decisions of a large number of people who are expected to be largely homogenous in terms of their cultural background and the levels at which particular category members are found. As these two predictions serve to distinguish prototype and exemplar theories, these are the lines of enquiry that this thesis will pick up.

Chapter 2 Experiment 1: A closed sentence-sorting study to test the representativity of linguists' intuitions about word senses

When linguists talk and write about polysemy, they tend to offer examples of polysemous words in action. Without such examples, an individual unfamiliar with polysemy might find the notion too abstract. The examples linguists offer might be constructed by the linguist themselves, or taken from a corpus. To demonstrate polysemy, i.e., to highlight the idea that words have distinct but related senses, linguists will offer examples of different senses. I did just that in Chapter 1. A linguist concerned with using examples to simply demonstrate the principle of polysemy need not necessarily trouble themselves with offering an explanation of how they determined that the examples they gave do indeed represent different senses. However, most academic treatments of polysemy are not concerned with offering a mere demonstration of the phenomenon, but are instead tasked with offering an account of some aspect of polysemy, or of a particular polysemous word. For example, a scholar might wish to observe the acquisition of the senses of a polysemous word (e.g., Rice, 2003), or they might want to study the processing of different senses (e.g., Foraker and Murphy, 2012). They may even wish to tackle the polysemy of a particular word head on, and offer an account of what the senses of that word are (e.g., Tyler and Evans, 2001; Brugman, 1981; Brugman and Lakoff, 2006 [1988]; of course, dictionaries are tasked with that objective also). In these cases, when we wish to offer an account of polysemy in any depth, and when that account depends upon proposing a set of senses for observation in acquisition, processing, and so on, we might wonder exactly how those senses were identified. This does, of course, question the integrity of the linguist's intuitions. Given their linguistics expertise in general, and likely expertise in meaning in particular, such questioning might be considered unfair. But the question remains: just how good are linguists' intuitions when it comes to word meaning?

This chapter aims to answer that question empirically. As stated in Chapter 1, and in line with current thinking in cognitive linguistics, this thesis assumes that polysemy is a case of linguistic categorisation. It is assumed in the field that polysemous words themselves are categories. At this point, I take an agnostic position on this issue, and do not reject the notion that it may be senses that are categories. In any case, it is assumed that, given their status as linguistic categories, polysemous words (or their senses) will behave like categories; i.e., it will be possible to organise examples of polysemous words into groups according to some categorisation criterion. In this case, the criterion of interest is the meaning of the polysemous word. Exploiting this assumption, I therefore use a categorisation task, operationalized as a sentence-sorting task, to assess whether the way I semantically categorise examples of the polysemous words *over, under, above* and *below* is systematically different from, or similar to the way these examples are categorised by other English speakers. If participants and I categorise the sentences in a similar way, we might conclude that this constitutes evidence that, in this case, expert intuitions about the senses of polysemous words *do* correspond to those held by other speakers. If it is not the case, it will cast doubt on the representativity of expert intuitions about word senses.

This study is situated in a broader literature that has interrogated the representativity and utility of expert intuitions about linguistic – and primary syntactic – phenomena. With that in mind, this chapter will open with an overview of relevant literature on the status of expert intuitions in linguistics in general, and in word senses and word meaning in particular. Following this literature review, I describe a set of closed sentence-sorting tasks carried out with a large group of predominantly naïve participants, which is designed to assess whether my intuitions about word senses correspond to theirs. The chapter closes with some concluding remarks on the implications of the findings, and identifies areas for further investigation.

Part 1: Literature review

This section will address the problematic status of expert intuitions in linguistic analysis. It will open by addressing the problem in general, before digging deeper to study the place of expert intuitions in the study of polysemy in particular. It will then discuss research which has moved away from reliance on intuitions, and identify some empirical approaches to studying polysemy and word senses.

2.1 Introspection as a methodology

Linguistic introspection, in which "conscious attention [is] directed by a language user to particular aspects of language as manifest in her own cognition" (Talmy, 2007, p. xii), is explicitly acknowledged as occupying a privileged position in the methodologies of choice in linguistics (Talmy 2007; Willems 2012; Schwarz-Friesel 2012). It is, for example, acknowledged that the development of syntactic theories depends heavily upon the use of introspection. As observed by Bradac, Martin, Elliott, and Tardy (1980), the traditional means for assessing whether a string is grammatical is simply a case of the researcher making a judgment. They argue that this method makes two rather shaky assumptions: that the researcher's stock of intuitions is the same as that held by all other native speakers of the language of interest, and that to judge a sentence is simply a case of checking one's intuitions and making a decision (p. 968). The first assumption seems reasonable when considering Spencer's (1973) observation that "a shared language is an important means of communication among members of a human society" (p. 83). This statement assumes that the grammar held by members of a given language community should be more or less the same. Accordingly, judgments about whether or not a sentence is grammatical can be given by any member of that community, since their shared grammar should result in convergent judgments. Since it has been demonstrated that different members of a language community do not necessarily acquire the same grammar (see section 3.1, where this issue is discussed in more detail), this assumption is unsound. Indeed, as Schütze (1996, p. 9) points out, these variations in interspeaker grammar may prove to be the source of interesting facts. Reliance on introspection as a methodology therefore rules out possibilities for discovering interesting findings about inter-speaker variation.

When we consider the use of syntacticians' (and, indeed, linguists in other fields) intuitions when developing theories, we must address the issue of their expert status. On the one hand, their training and intense engagement with particular elements of syntax as both an individual working with it, and as a reader of literature about it, could be argued to render them best qualified to judge whether or not a particular sentence is grammatical. But on the other hand, is this extreme familiarity with a given construction, for example, a problem? Snyder (2000) and Spencer (1973) argue that it might be, and that an example of a particular construction initially

deemed ungrammatical might, with exposure (which Snyder has observed need not be prolonged), come to be judged to be acceptable. Accordingly, a syntactician whose research is concerned with a particular construction might ultimately decide that a particular example is acceptable when someone unfamiliar with the construction – for example, a naïve speaker – might deem it ungrammatical.

In an energetic debate, studies have found evidence indicating divergence between predictions of syntactic theory or linguists' grammaticality judgments and non-linguists' grammaticality judgments (Bradac et al., 1980; Dąbrowska, 2010; Gordon and Hendrick, 1997; Ross, 1979; Schütze, 1996; Spencer, 1973); elsewhere, evidence is offered indicating that where there is divergence, it is on a minute scale (Sprouse and Almeida, 2012; Sprouse, Schütze, and Almeida, 2013). Given that some studies show that naïve subjects are found to disagree with linguists' acceptability judgments, and given that the task of asking a sample of non-linguists to contribute their own judgments is not an onerous one (Gibson and Fedorenko, 2013), it seems reasonable to recommend that theories are grounded in data collected from a much larger and more representative sample of speakers than the author and/or a handful of his colleagues or students. On the grounds that research has revealed divergences between linguists' and non-linguists' judgments – discoveries that are not necessarily undermined by independent data to the contrary – it seems that Labov was correct when he argued "that linguists cannot ... produce theory and data at the same time" (1972, p. 199). Further, since contributions from non-linguists are easily gathered, acquiring unbiased data to check a theory is an inexpensive but highly valuable enterprise.

The use of introspection and reliance on an author's own intuitions as the sole methodology underpinning a particular part of research has a number of important flaws, and it is important that any study adopting this methodology acknowledges its limitations. As we have already seen, the use of introspection to describe a particular linguistic phenomenon has been demonstrated to be problematic, with the author's intuitions not reliably coinciding with those of non-linguists. The situation for introspection is no better when the goal of a particular research project is to explain a particular phenomenon. As has been noted by Gibbs (2006), Schwarz-Friesel (2012), and Talmy (2007), conscious access to the unconscious processes understood to

constitute language is not possible. This conclusion is analogous to Miller's (1962, p. 56) argument that "[i]t is the result of thinking, not the process of thinking, that appears spontaneously in consciousness" and is therefore available for introspective analysis. Gibbs suggests that our inability to access the mental processes that underpin language is due to inconsistencies in the characteristics of consciousness and unconsciousness. As Talmy (2007) observes in a chapter that gives rather a lot of credence to introspection as a methodology, the scope of outputs generated by a trawl of one's own intuitions does not necessarily reflect the possible range of outputs. To take his example, and one which is relevant to this research, an introspective investigation into the senses of a polysemous word does not yield "a full connected set [...] though it does reveal a few [senses]" (p. xiv). This admission undermines exclusively intuition-based analyses of polysemous words.

An additional issue that should concern those basing conclusions on intuitions is the possibility that an author's bias may have distorted the intuitions that were reported. Bias in this context has two forms: theoretical bias, and bias towards or away from particular "data". Dąbrowska's (2010) study of grammaticality judgments of questions with long-distance dependencies explicitly examined, in addition to (dis)agreement between linguist and naïve judges, the extent to which linguists with different theoretical persuasions agreed with one another. On the whole linguists identifying as cognitive-functionalists and generativists tended to reach fairly similar decisions, with only one significant exception. Generally, though, the former group's responses were closer to the judgments made by non-linguists. Dąbrowska concludes that more research is necessary to understand the relationship between theoretical commitments and grammaticality judgments.

2.2 Introspection, intuitions and polysemy

2.2.1 The status of introspection in the study of polysemous words

The second type of bias relevant in critiques of introspection as a methodology, and which is particularly relevant to the present study, is an author's bias towards or against particular "data" generated by a sweep of one's store of linguistic phenomena. For example, in a mental search for examples of a polysemous word one wishes to examine, an author may be inclined towards examples which provide "evidence" in support of a particular sense. An author may similarly be disinclined to

acknowledge more troublesome examples that would be difficult to ascribe to a particular sense. A credible means of acquiring examples of a particular sense is through extraction of corpus hits. Concordance lines extracted at random from the hits generated in a search of a particular word frequently reveal examples of senses one would expect, but they have the perfectly reasonable but nonetheless frustrating habit of either giving you lots of examples of a small set of senses, or a few examples that prove resistant to easy categorisation. The analysis of concordance lines therefore represents a significant improvement over the reliance on an author's set of auto-generated "examples", simply because they reveal the truth about how a particular word is used, whether that truth is convenient or not.

Of particular relevance to the study of polysemy is the problematic effect of semantic satiation, as crystallised by James (1962) and studied extensively since. Semantic satiation is, briefly, that strange effect in which repeated visual or aural exposure to a word results in its temporary loss of meaning. Studies of polysemy tend to focus on a single or at most a small number of words. While this has the benefit of providing focus on the nuances of a particular word or a small selection thereof, the degree of focus required to fully explore these nuances puts the author at risk of falling victim to semantic satiation. It has been my own experience, in the course of this study, that I find myself wondering exactly what *above* means: there have been moments when all I see is a string of letters. When an author has committed his or herself to the task of teasing out the senses of a particular word, he or she must fight the effects of semantic satiation to produce a credible analysis. This presents a problem to any author investigating a polysemous word, but it is reasonable to speculate that it will be particularly problematic to authors attempting to discern sense boundaries in a set of examples of the target word.

It must be acknowledged that this particular issue – of semantic satiation when faced with a set of instance of a given word – may affect the individuals participating in the present research. This, coupled with Stefanowitsch's (2011) concerns over whether or not the collection of intuitions from naive subjects is any clearer than the traditional introspection methodology, problematises the approach to identifying sense distinctions that this study takes. However, the approach is justified on two bases. First, in response to possible criticism over the possible effects of semantic

satiation on the participating subjects, the studies reported here has been designed to mitigate these effects, by presenting subjects with only a limited number of example sentences. This decision was made on the basis of recommendations from authors such as Miller (1971), who proposes that sorting a set of 100 stimuli is a manageable task – though others, such as Baker (1999) have used considerably more. It is not the aim of this study to identify all of the senses of the target words, but instead to understand certain aspects of their psychological status. For this reason, presentation of a limited number of sentences is defensible. Second, while Stefanowitsch's concerns are valid, and should be acknowledged in any study with a methodology similar to the one described here, as Tuggy (1999, p. 358) notes, "[i]f many speakers of a language coincide in an intuition regarding meaning (e.g., that a particular U1 and U2 can be distinguished, or that they are the same meaning, or both), that intuition should be accorded a high degree of credence." Further, Tuggy argues that when intersubjective consensus differs from the authors' own intuitions, "it has a very strong claim to be objective" (p. 358). Labov (1972, p. 106) also acknowledges the importance of seeking intersubject agreement with an author's intuitions. While the analysis of a large sample of speakers' intuitions can serve to check an authors' intuitions, they are also useful when *not* compared against an author's own impressions: as Gibbs (2006) observes, they can be used to uncover general patterns, and to identify individual quirks and differences. Moreover, results of categorisation tasks – which have significant precedent in the study of word meaning in and beyond cognitive linguistics – suggest decisions which are taken as *indications* of word senses. In this way, they allow the researcher to gather implicit, and so perhaps more reliable, conclusions.

Talmy (2007) presents a defence of the role of introspection in the study of meaning. He goes as far as to say that "introspection has the advantage over other methodologies in seemingly being the only one able to access [meaning] directly" (p. xiii). He argues that this bold statement is justified on the basis that meaning and introspection are both consciousness phenomena. Without offering any evidence for either of these claims, it is difficult to judge whether this is a logical conclusion. A number of questions arise out of Talmy's claims. If meaning is a consciousness phenomenon, and is open to the conscious process of introspection, is the comprehensive meaning potential of a word available at once? In other words, are all

senses/exemplars of a given word available for simultaneous introspective analysis? Given his later comment that not all senses of a particular word come to mind under an introspective analysis, it seems that the answer to this question is likely to be no. If the answer *is* no, how can one commit to describing the full range of senses of a polysemous word on the basis of introspection alone? There is certainly a role to be played by introspection in the study of language, such as in the formulation of hypotheses to be tested, and research questions to be answered. However, his statement attesting that introspection is the best means of understanding meaning overplays its power, and overlooks its weaknesses. Amidst his bold statements can be found a more balanced account of the role of introspection, and he acknowledges its limited application to other aspects of linguistic research. It is fortunate that we find ourselves in a period in which the power and necessity of experimental and other empirical work is being acknowledged and embraced (Arppe & Järvikivi 2007).

2.2.2 Intuitions in use: Cognitive linguistics

Problems surrounding authors' reliance on introspection and intuitions in the development of theories about and observations of polysemy have been acknowledged. For example, Sandra and Rice (1995) detail a number of issues including the lack of a "clear-cut methodology … for making distinctions between prepositional usages" [p. 90] for identifying which senses are real and constitute meaningful distinctions, and raise concerns regarding the granularity of sense distinctions proposed. Despite this, introspection as a methodology persists in polysemy research. In a proposal which was intended to respond to Sandra and Rice's criticisms over the lack of rigorous principles underpinning sense identification, Tyler and Evans (2001) put forward the Principled Polysemy model, which aims to provide a principled methodology for identifying the senses of a given word, and for identifying which of those senses is the "protosense" from which all other senses extend. While a principled and well-articulated approach to sense distinction is welcome, and while publishing the criteria for making these discoveries in principle allows a particular study to be replicated by other researchers, since the majority of the criteria require mental processing by the individual researcher, this proposal remains lodged in the introspective tradition. To date, no research in which a different scholar replicates a Principled Polysemy study

of a word already studied under this protocol has been undertaken. For that reason, Tyler and Evans' claim as to the replicability of the procedure (2001, p. 731) remains that – a claim – rather than a tested reality.

Since publication of Tyler and Evans' seminal piece on *over*, other studies have continued to use the introspective methodology. Mahpeykar and Tyler (2015) have used the Principled Polysemy method in an analysis of the phrasal verbs *with, up* and *out*, and in an earlier study of the Farsi preposition *be* (Mahpeykar and Tyler, 2011). The results of Tyler and Evans' (2001) study of *over* is used in Ostermann's (2014) efforts to build a dictionary entry for the same word that captures the relations between its senses, and Masi's (2010) study of *over* and its Italian counterpart *sopra*. The Principled Polysemy method was used again by Evans and Tyler (2004) to distinguish between the senses of the preposition *in*. Beyond research adopting Principled Polysemy, Kishner and Gibbs' (1996) study of *just* studies eight senses of this word, six of which were identified by another author, and two which they identified. No explanation was given for how these senses were identified, suggesting that they were distinguished on the basis of their intuitions.

2.2.3 Intuitions in use: Computational linguistics

Expert intuitions are at the heart of efforts to develop productive automated word sense disambiguation (WSD) algorithms. These algorithms make use of sense-tagged corpora of the target language. Those senses are traditionally matched to instances of each word by trained lexicographers – which I will, following convention, refer to here and in other sections concerning WSD as *annotators* – using a resource such as a dictionary or WordNet. At least two annotators tag each word, following their intuitions, and their decisions are checked for agreement. Sense-tagged datasets are used as "gold standards" against which the performance of a WSD system can be evaluated. Kilgarriff (1998) describes the necessity of a gold standard dataset, and the difficulties inherent in creating one. He notes that "The pervasive worry in preparing the dataset [for a gold standard] is that it will not meet adequate standards of replicability: that is, if two people tag the same text, they will all too frequently assign different tags to the corpus instance." (p. 16). Indeed, it has been demonstrated that even trained annotators do not necessarily reach consensus with each other (Passonneau et al. 2010).

2.2.4 Empirical approaches to identifying word senses

A range of empirical approaches to the study of polysemy and sense distinction are being taken, such as behavioral profiles (Gries and Divjak, 2009; cf. Berez and Gries, 2008; Gries, 2006), in which concordance lines are analysed for a vast range of linguistic features and, using statistical analyses such as clustering, groups of usages that share similar features are isolated, which are taken suggest distinct senses. In addition, Srinivasan and Rabagliati (2015) have used tests to see whether patterns of polysemy, such as container for contents (e.g., *bowl*), and object for representational content (e.g., *book*), which result in what the authors argue to be distinct senses of English words, are evident in other languages. Durkin and Manning (1989) identified the senses of polysemous words by asking speakers to assign a meaning of their choosing to a stimulus ambiguous (either homonymous or polysemous) word. Rather than presenting subjects with the stimulus word being used in a particular manner, subjects were asked to provide a single statement of the meaning of the target word presented in isolation, specifically the first that came to mind. While there was some crossover in the meanings given by different subjects, and while a particular meaning of a given word was provided more frequently than others, at least two different senses were identified for all 175 stimulus words.

In his examination of polysemous word senses, Baker (1999) uses four psycholinguistic investigations of subjects' responses to the word *see*: an open sentence-sorting task, a closed sentence-sorting task, a lexical decision task and a categorical judgment task. An open sorting task allows researchers to understand what distinctions subjects make when not constrained or distracted by the provision of pre-set categories. Subjects sorted examples of *see* extracted from corpora, and a small number that had been constructed by the author. Over the course of three sets of experiments, subjects sorted sentences reflecting, in the eyes of the author, 23 senses. Instructed to attend solely to the meaning of the target word, subjects were asked to sort the examples into groups according to common meaning and, at the end, either give each group a short definition (in the case of experiment one), or, in the case of experiments two and three, identify which example in the pile was most representative of the meaning reflected in that pile. Baker's closed sorting tasks entail presenting subjects with a series of sentences and categories consisting of the senses of *see* that the author had identified, and asking them to allocate each stimulus

sentence to one of the given senses. Lexical decision and categorical judgment tasks were used to understand how responses vary between off-line and on-line tasks. In the lexical decision task, after being shown an example sentence seen previously in the sorting tasks, subjects heard a probe (non-)word, and were asked to decide whether or not it was a word. In the categorical judgment task, the same protocol was used, but subjects were asked to decide whether or not the probe word was an instance of the sense primed by the example sentence.

Baker's comprehensive approach to understanding polysemous word senses is rather unusual in the literature on polysemy. While the studies mentioned here, in addition to those using measures such as sorting and timed tasks (cf. Rice, Sandra, and Vanrespaille, 1999; Rice, 1996; Sandra and Rice, 1995), suggest that interest in testing our intuitions about polysemy and word senses is growing, the scale of Baker's investigation is exceptional. This particular work was, however, concerned with establishing exactly what the senses of *see* are, and so we might attribute this unusual degree of conscientiousness to the study's ultimate goal. Indeed, while Baker noted that there was variation in how well participants agreed with his sense distinctions (p. 147), he does not proceed to discuss this interesting outcome in any detail.

2.3 Expert intuitions about word senses: conclusions

This section has demonstrated a conflict between approaches to linguistics research: on the one hand, there is a growing argument that linguists' intuitions in isolation should not be understood as data, or as evidence in support of any particular conclusion. On the other hand, and in spite of methodological advancements demonstrating that intuitions need not be the sole or main source of data, intuition-led approaches to the study of polysemy occupy privileged positions in the cognitive linguistic canon. While principled approaches to the delineation of word senses have been put forward (Tyler and Evans, 2001), and while this approach adopts an explicit methodology, it remains a subjective approach that has thus far not been repeated by another author in any published analysis of *over*, the word Tyler and Evans studied. The extent to whether their decision principles produce replicable outcomes is therefore unknown.

Arriving at a set of sense distinctions in an objective manner is highly desirable. If we are able to isolate the senses of a given word, we may then begin the theoretically interesting pursuit of an account of whether, and how, those senses are related to each other. At a practical level, identifying the senses of a given word is a necessary step in dictionary writing and when developing sense inventories used to train word sense disambiguation algorithms. It is vital that the steps taken to achieve these objectives can be replicated by other researchers in a way that produces the same results. Evidence about whether or not linguists' intuitions about the senses of a given polysemous word correspond to those of other speakers is lacking. In the face of evidence of divergence between expert and naïve intuitions about other linguistic phenomena (Bradac et al., 1980; Dąbrowska, 2010; Gordon and Hendrick, 1997; Ross, 1979; Schütze, 1996; Spencer, 1973), it is necessary to ask how representative – and consequently how useful – expert intuitions about word meaning are.

Part 2: Investigation

The second part of this chapter describes a closed sentence-sorting task that aims to establish whether the senses that I, as a trained linguist, find meaningful are meaningful in the minds of other speakers. The structure of this part of the chapter is as follows. Based on the literature review presented above, I identify the gaps in knowledge that I wish to address, and specify the aims of the study. I then present some of the senses of the words *over, under, above* and *below* that I have identified, guided by my intuitions. I then move to report the results of a sentence-sorting task, analogous to categorisation experiments used in non-linguistic categorisation research and WSD tagging exercises. I first undertake a qualitative analysis of how participants as a group tended to categorise sentences, to assess whether any patterns emerge. Afterwards, I use Cohen's kappa to statistically measure how well individual participants and I agree about how the stimuli should be sorted. This part of the chapter closes with some concluding remarks on the implications of the findings, and how they can be tested further.

2.4 Aims

The literature review in Part 1 of this chapter revealed that, in spite of growing awareness that a linguist's intuitions about a particular linguistic phenomenon may

not correspond to those held by other speakers, intuitions continue to occupy a privileged position in the methodology of choice in the study of polysemy. While experimental research on the representativity of linguists' intuitions about syntactic phenomena has been widely reported, equivalent interrogation of the status of linguists' intuitions about word senses has not.

I therefore aim to offer an original contribution to knowledge intended to fill this gap. Specifically, I aim to carry out an empirical investigation of how well naïve speakers' and other linguists' intuitions about word senses correspond to my own. The findings of this study are intended to shed further light on the utility of expert intuitions, and therefore add to existing literature on this issue to expand the focus beyond syntax and into semantics.

2.5 The senses of over, under, above and below: A linguist's view

The flexibility of use of these four polysemous words presents a considerable challenge to the linguist tasked with disentangling their uses, and classifying them into sense groups. This is particularly the case when these uses are drawn from corpora; as Berez and Gries (2008) observe, corpus outputs can be surprising, returning a diverse sample that may be trickier to categorise than examples constructed by a linguist. Constructed examples may be produced to support the proposition that a particular sense exists, thus rendering that sense a self-fulfilling prophecy. The diverse sample that can be collected from a corpus of real usage allows the linguist to establish objectively how a particular word is used, and may well reveal uses that an introspective trawl of examples of the target word would not have returned. While using corpora as the basis of a usage classification task is therefore rather more challenging than using constructed examples, it is a more rigorous – not to mention replicable – procedure that may indeed return more interesting results than introspection can.

Tables 1 to 4 below show the sense categorisation decisions I made when sorting examples of the four target words. Following the tables, the sense that I judge to be used in each group is described.

Table 1 Stimuli for *over* sorting task, categorised into senses according to my intuitions

Sentence	Group label	Sense
There was a wrangle OVER a proposal to adopt a law.	They spent two weeks squabbling OVER the issue.	ABOUT
I puzzled OVER this.	They spent two weeks squabbling OVER the issue.	
Let's not fight OVER it.	They spent two weeks squabbling OVER the issue.	
Clashes also occurred OVER trapping rights.	They spent two weeks squabbling OVER the issue.	
We had some discussions OVER where to place the boundaries.	They spent two weeks squabbling OVER the issue.	
We'd fall out OVER stupid things and not speak to each other.	They spent two weeks squabbling OVER the issue.	
Can you just run it OVER the road?	He sped up as he drove OVER the bridge.	A-B MOVEMENT (NO ARC)
The cops pulled me OVER.	He sped up as he drove OVER the bridge.	
I ran OVER the bridge.	He sped up as he drove OVER the bridge.	
Sarah's come OVER the road Daddy.	He sped up as he drove OVER the bridge.	
The plane flew OVER the city.	He sped up as he drove OVER the bridge.	
He walked slowly OVER the zebra-crossing.	He sped up as he drove OVER the bridge.	
Jump OVER the other one.	John was shot as he climbed OVER the Wall.	ARC
Herons seem to be incapable of stepping OVER the deterrent.	John was shot as he climbed OVER the Wall.	
I go OVER the handlebars.	John was shot as he climbed OVER the Wall.	
The quick brown fox jumped OVER the lazy dog.	John was shot as he climbed OVER the Wall.	
He refused to return the balls kicked OVER his fence.	John was shot as he climbed OVER the Wall.	
They keep slinging their towels OVER the bedroom door.	John was shot as he climbed OVER the Wall.	
He is handing OVER his presidency.	Meridian TV are taking OVER from TVS.	TRANSFER
That's half the reason that Brian Tolbrook took OVER at Tettron.	Meridian TV are taking OVER from TVS.	
He took OVER the printing business.	Meridian TV are taking OVER from TVS.	
I'll take OVER the primary agenda.	Meridian TV are taking OVER from TVS.	
I can't hand OVER a long barrelled weapon to that officer.	Meridian TV are taking OVER from TVS.	
The plaintiff handed OVER to Samuel Revill the first note.	Meridian TV are taking OVER from TVS.	
I rushed out before the show was OVER.	She agreed to meet Tessa when the visit was OVER.	COMPLETION
I can't believe the weekend is OVER already!	She agreed to meet Tessa when the visit was OVER.	
His wrestling days are not OVER yet.	She agreed to meet Tessa when the visit was OVER.	
They think it's all OVER…it is now!	She agreed to meet Tessa when the visit was OVER.	
She wished the party was OVER.	She agreed to meet Tessa when the visit was OVER.	
We hope Sunderland will go up once the game is OVER.	She agreed to meet Tessa when the visit was OVER.	
Bring pelmet fabric OVER, and press with an iron to bond together.	He turned OVER the pages of the notebook.	FLIP
I turn it OVER.	He turned OVER the pages of the notebook.	
The printed sheets are turned OVER on the long axis.	He turned OVER the pages of the notebook.	
Turn that steak OVER, it's burning!	He turned OVER the pages of the notebook.	
Yeah can you turn that OVER please.	He turned OVER the pages of the notebook.	
He saw a car flip OVER and land upside down in a hedge.	He turned OVER the pages of the notebook.	

Table 2 Stimuli for *under* sorting task, categorised into senses according to my intuitions

Sentence	Group label	Sense
I'm wearing a vest UNDER this shirt.	He's wearing pyjamas UNDER his jacket	HORIZONTAL RELATIONSHIP
I'm a bit hot UNDER the collar.	He's wearing pyjamas UNDER his jacket	
Should I wear a jumper UNDER this coat?	He's wearing pyjamas UNDER his jacket	
They got UNDER cover of the walls of the fortresses.	He's wearing pyjamas UNDER his jacket	
Rinse the dish UNDER running water.	He's wearing pyjamas UNDER his jacket	
They found more wallpaper UNDER the layer they'd removed.	He's wearing pyjamas UNDER his jacket	
You can emigrate to Britain UNDER limited criteria.	It's allowed UNDER the terms of the policy.	ACCORDING TO
He will be committed UNDER the mental health act.	It's allowed UNDER the terms of the policy.	
It will be contested UNDER the Corrupt Practices Act.	It's allowed UNDER the terms of the policy.	
They will file an explanation UNDER Article 51 of the UN Charter.	It's allowed UNDER the terms of the policy.	
UNDER the new regulations, the students must sign in each week.	It's allowed UNDER the terms of the policy.	
The fridge can be exchanged UNDER the returns policy.	It's allowed UNDER the terms of the policy.	
Remember their hospital is UNDER threat.	We kept the place UNDER observation.	SUBJECT TO
This is something which is very much UNDER attack.	We kept the place UNDER observation.	
They deny they were UNDER any duty to offer any advice.	We kept the place UNDER observation.	
We're UNDER real pressure at the moment.	We kept the place UNDER observation.	
Your application is now UNDER review.	We kept the place UNDER observation.	
The question of intercommunion is UNDER discussion.	We kept the place UNDER observation.	
I'm frying the bread UNDER there.	He's standing UNDER the mistletoe.	VERTICAL RELATIONSHIP, NO CONTACT
I'm hiding UNDER your bed.	He's standing UNDER the mistletoe.	
The Troll seldom came out from UNDER the bridge.	He's standing UNDER the mistletoe.	
She kicked him UNDER the table.	He's standing UNDER the mistletoe.	
They lay UNDER the stars.	He's standing UNDER the mistletoe.	
Cover dark circles UNDER your eyes with concealer.	He's standing UNDER the mistletoe.	
I'm sleeping UNDER my cover.	I hid the crumbs UNDER the rug	VERTICAL RELATIONSHIP WITH CONTACT
The growing roots UNDER the path had cracked the tarmac.	I hid the crumbs UNDER the rug	
They put material UNDER the carpet to make it more comfortable.	I hid the crumbs UNDER the rug	
They looked where others wouldn't think to - UNDER dark leaves.	I hid the crumbs UNDER the rug	
They'd been hiding people UNDER the floorboards.	I hid the crumbs UNDER the rug	
He felt warm UNDER the blanket.	I hid the crumbs UNDER the rug	
We'll be taxed UNDER the Conservatives.	We're UNDER new management.	UNDER THE CONTROL OR AUTHORITY OF
I'm working UNDER the direction of the area manager.	We're UNDER new management.	
250 Garrimperos work UNDER him.	We're UNDER new management.	
She's got a whole team working UNDER her now.	We're UNDER new management.	
He served in 102 Battalion UNDER the South Africans.	We're UNDER new management.	
The Act was introduced UNDER the last President.	We're UNDER new management.	

Table 3 Stimuli for *above* sorting task, categorised into senses according to my intuitions

Sentence	Group label	Sense
They think they're ABOVE work like this.	Working conditions are ABOVE average.	BETTER THAN
Are they good, ABOVE average, or below average?	Working conditions are ABOVE average.	
It was either ABOVE average or below average.	Working conditions are ABOVE average.	
The Renault 5 was just ABOVE banger status.	Working conditions are ABOVE average.	
She's not ABOVE silly gossip.	Working conditions are ABOVE average.	
I'm ABOVE all that petty business.	Working conditions are ABOVE average.	
Anything ABOVE zero degrees and the ice will melt	You should bid ABOVE the asking price.	MORE THAN
He has a surplus of votes over and ABOVE the quota.	You should bid ABOVE the asking price.	
It was 40% ABOVE £150.	You should bid ABOVE the asking price.	
The new estimate is 95,000 ABOVE the original estimate.	You should bid ABOVE the asking price.	
The price of fuel has jumped ABOVE $4 a gallon.	You should bid ABOVE the asking price.	
Train fares have risen ABOVE inflation.	You should bid ABOVE the asking price.	
There was a faint bruise ABOVE her eyebrow.	They live in the flat ABOVE the shop.	VERTICAL RELATIONSHIP
I've hung some mistletoe ABOVE the doorway.	They live in the flat ABOVE the shop.	
The dictionaries are ABOVE the history books.	They live in the flat ABOVE the shop.	
The plane was cruising ABOVE the clouds.	They live in the flat ABOVE the shop.	
The shelf is fixed to the wall ABOVE the radiator.	They live in the flat ABOVE the shop.	
The stars ABOVE were partly obscured by clouds.	They live in the flat ABOVE the shop.	
The comments (listed ABOVE) are worrying.	There are several issues, as listed ABOVE.	TEXT USES
As described ABOVE, this uses a new operating system.	There are several issues, as listed ABOVE.	
It was refused for the ABOVE reasons.	There are several issues, as listed ABOVE.	
The process, described ABOVE, is clear to all.	There are several issues, as listed ABOVE.	
All of the ABOVE laws have been passed in the last ten years.	There are several issues, as listed ABOVE.	
The ABOVE constraints are not seen as insurmountable.	There are several issues, as listed ABOVE.	
This site is elevated ABOVE the road.	We camped ABOVE the valley floor.	VANTAGE
Glastonbury Tor towers ABOVE the Somerset Levels.	We camped ABOVE the valley floor.	
It was built on the hill, just ABOVE the station.	We camped ABOVE the valley floor.	
We had a great view from the cliff ABOVE the cove.	We camped ABOVE the valley floor.	
The town is 200m ABOVE sea-level.	We camped ABOVE the valley floor.	
We were observed from the window ABOVE.	We camped ABOVE the valley floor.	
I used to be his boss, but he works ABOVE me now.	There's nobody ABOVE them in the hierarchy	HIERARCHY
Now that I work ABOVE her, we don't talk so much.	There's nobody ABOVE them in the hierarchy	
The orders came from ABOVE.	There's nobody ABOVE them in the hierarchy	
There is a level of executives ABOVE the vice president level.	There's nobody ABOVE them in the hierarchy	
ABOVE private soldiers there are three types of officer.	There's nobody ABOVE them in the hierarchy	
We're under serious pressure from ABOVE to get the job done.	There's nobody ABOVE them in the hierarchy	

Table 4 Stimuli for *below* task, categorised into senses according to my intuitions

Sentence	Group label	Sense
Congress is somewhere BELOW cockroaches and traffic jams in Americans' esteem.	They were of BELOW average ability.	WORSE THAN
We must set standards of achievement BELOW which they must not fall.	They were of BELOW average ability.	
You wouldn't be doing the job if you were BELOW that level.	They were of BELOW average ability.	
Paul had performed BELOW expectation.	They were of BELOW average ability.	
He performed BELOW par last time.	They were of BELOW average ability.	
He is an unenthusiastic and BELOW average soldier.	They were of BELOW average ability.	
Tabith stood BELOW, watching him.	From the window I saw the field BELOW.	VANTAGE
BELOW the front windows the extension was divided into two sections.	From the window I saw the field BELOW.	
They established an iron foundry in the valley BELOW the church in 1790.	From the window I saw the field BELOW.	
The walk provides wonderful views of Mallerstand BELOW.	From the window I saw the field BELOW.	
Your mates are down BELOW, watching you.	From the window I saw the field BELOW.	
I called out to the people on the beach BELOW, but they didn't hear me.	From the window I saw the field BELOW.	
The people in the flat BELOW wouldn't stop shouting.	Hundreds of pipes run BELOW the city.	UNDERNEATH (3D)
Instead of being up high the box was down BELOW.	Hundreds of pipes run BELOW the city.	
We dredged BELOW the mud at the bottom of the river.	Hundreds of pipes run BELOW the city.	
When we got BELOW the next layer the concentrations became stronger.	Hundreds of pipes run BELOW the city.	
The campus was shrinking BELOW me into a collection of children's play houses.	Hundreds of pipes run BELOW the city.	
The crocodile sank BELOW the surface.	Hundreds of pipes run BELOW the city.	
She had a mole just BELOW her right eye.	There was a large scratch BELOW the driver's window.	LOWER THAN (2D)
Don't paint BELOW the windowsill.	There was a large scratch BELOW the driver's window.	
BELOW the crags a well-built tunnel could be seen.	There was a large scratch BELOW the driver's window.	
I pinned my name badge BELOW the logo on my tshirt.	There was a large scratch BELOW the driver's window.	
There are two iron rings on the wall BELOW the painting.	There was a large scratch BELOW the driver's window.	
The sleeves gradually get tighter and end BELOW the elbow.	There was a large scratch BELOW the driver's window.	
The loss is a little BELOW £3,200.	The group has been trading BELOW budgeted levels.	LESS THAN
There's no level BELOW which the wages may not fall.	The group has been trading BELOW budgeted levels.	
We brought in forty million pounds BELOW the target amount.	The group has been trading BELOW budgeted levels.	
He set a price BELOW the existing supplier's 1994 prices.	The group has been trading BELOW budgeted levels.	
There is a £20 surcharge on orders BELOW £50.	The group has been trading BELOW budgeted levels.	
The sales value was well BELOW target.	The group has been trading BELOW budgeted levels.	
Fill in BELOW all the tasks that you do in a typical day.	Tickets are available by completing the form BELOW.	TEXT USES
Look at the sentence BELOW, what does it say?	Tickets are available by completing the form BELOW.	
Give us your fun verdict by dialling the numbers BELOW.	Tickets are available by completing the form BELOW.	
Serve with Sharp sauce (see BELOW).	Tickets are available by completing the form BELOW.	
In the situations listed BELOW identify what your information needs would be.	Tickets are available by completing the form BELOW.	
It will be argued BELOW that economic reconstruction was a success.	Tickets are available by completing the form BELOW.	

2.5.1 Distinction procedure

I followed an intuition-based approach to distinguishing between the example sentences in the four tables above, classifying examples according to whether I judged that they used the same or a different sense of the target word. A large, random sample was extracted and each was annotated to describe, or paraphrase, what I judged to be the underlying sense used in that context. Once annotated, the sample was sorted according to the description/paraphrase, thus revealing initial groups of sentences. By comparing these grouped sentences alongside each other, I was able to verify whether each did indeed capture the same sense of the target word, whether any had been mis-annotated, or whether I judged that there was a sufficiently meaningful distinction within the group to license splitting it further. Six examples of six senses of each word were selected to be included in the tasks. This was a conscious decision: it ensured that the task was of a manageable scale, which should maximise task completion, while minimising fatigue and semantic satiation effects. The senses are defined in the following sections.

2.5.2 Over

2.5.2.1 ABOUT

Each of these sentences describes the effect of a triggering stimulus. The group does not distinguish between the physicality of the action that takes place as a consequence, and therefore captures physical action, as in *Let's not fight over it*, and non-physical action, as in *I puzzled over this*. The figure (i.e., in these examples, *this* and *it*) in each sentence thus acts as both the trigger for the action and consequently the object of the action.

2.5.2.2 A-B MOVEMENT (NO ARC)

Each sentence in this group describes the movement of an object or person from one location to another. The majority of the stimuli capture a path from a side of one location to another. For example, *Sarah's come over the road Daddy* describes the movement by *Sarah* from one side of the street to the side on which the speaker is located. The group does not specify for degree of contact or distance between figure and ground, as indicated by the presence of the sentences *The plane flew over the city*, and *He walked slowly over the zebra-crossing*. Further, it does not specify that exemplars must describe a horizontal path as in Figure 6.

2.5.1 Distinction procedure

I followed an intuition-based approach to distinguishing between the example sentences in the four tables above, classifying examples according to whether I judged that they used the same or a different sense of the target word. A large, random sample was extracted and each was annotated to describe, or paraphrase, what I judged to be the underlying sense used in that context. Once annotated, the sample was sorted according to the description/paraphrase, thus revealing initial groups of sentences. By comparing these grouped sentences alongside each other, I was able to verify whether each did indeed capture the same sense of the target word, whether any had been mis-annotated, or whether I judged that there was a sufficiently meaningful distinction within the group to license splitting it further. Six examples of six senses of each word were selected to be included in the tasks. This was a conscious decision: it ensured that the task was of a manageable scale, which should maximise task completion, while minimising fatigue and semantic satiation effects. The senses are defined in the following sections.

2.5.2 Over

2.5.2.1 ABOUT

Each of these sentences describes the effect of a triggering stimulus. The group does not distinguish between the physicality of the action that takes place as a consequence, and therefore captures physical action, as in *Let's not fight over it*, and non-physical action, as in *I puzzled over this*. The figure (i.e., in these examples, *this* and *it*) in each sentence thus acts as both the trigger for the action and consequently the object of the action.

2.5.2.2 A-B MOVEMENT (NO ARC)

Each sentence in this group describes the movement of an object or person from one location to another. The majority of the stimuli capture a path from a side of one location to another. For example, *Sarah's come over the road Daddy* describes the movement by *Sarah* from one side of the street to the side on which the speaker is located. The group does not specify for degree of contact or distance between figure and ground, as indicated by the presence of the sentences *The plane flew over the city*, and *He walked slowly over the zebra-crossing*. Further, it does not specify that exemplars must describe a horizontal path as in Figure 6.

Figure 6 Horizontal, linear path

As inclusion of the example *The cops pulled me over* suggests, the group accommodates paths which can be diagonal or otherwise non-horizontal in nature, as illustrated in Figure 7 and Figure 8.

Figure 7 Diagonal, linear path

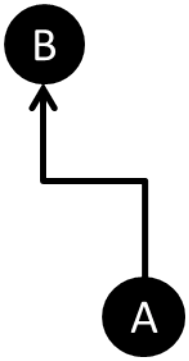


Figure 8 Non-linear path

2.5.2.3 ARC

Whereas the previous group collectively described relatively flat motion from one point to another, the sentences in the ARC category are further specified to describe the movement from A to B in which the figure moves up and over an obstacle of varying height; from a *lazy dog* to a *fence*; in brief, the trajectory is arc-shaped. Perhaps the least exemplary of the six sentences is *They keep slinging their towels over the bedroom door*. In this case, the sentence describes the movement of one part of the figure (*towels*) over an obstacle (*the bedroom door*), finishing on the other side of the obstacle (point B). However, whereas in the other examples the entire figure ends up at point B, parts of the figure in this sentence remain at points A *and* B. In this case, the shape of the figure at the end of the motion matches the shape of the trajectory of the figures in the other sentences.

2.5.2.4 TRANSFER

Each sentence in this group captures the transfer of ownership or responsibility for a physical or abstract entity from one person to another. The group does not make a distinction between physical transfer, such as that captured by *The plaintiff handed over to Samuel Revill the first note* and abstract transfer, as exemplified by *I'll take over the primary agenda*.

2.5.2.5 COMPLETION

This group of non-spatial examples of *over* collectively capture a temporal sense. Specifically, they describe the end of some temporal event: a *weekend*, a *show*, or an unspecified event as in *They think it's all over…it is now!*.

2.5.2.6 FLIP

These spatial sentences describe the inversion movement of a figure. Figure 9 illustrates this and is discussed afterwards.

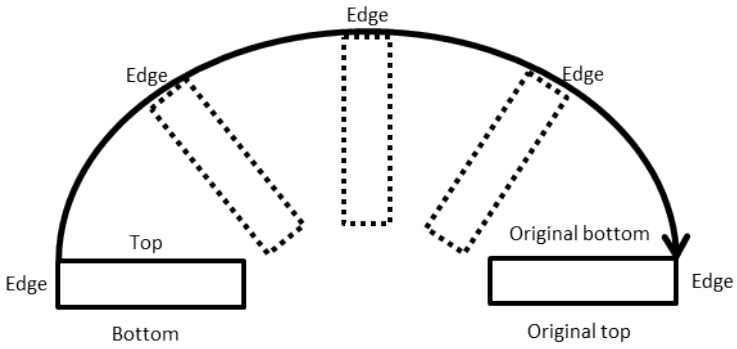


Figure 9 Schematic representation of FLIP, in which the path of the edge of the figure object corresponds to the shape of the trajectory underlying the ARC sense

As the illustration suggests, the sentences describe the inversion of an object such that what was its top part at an earlier point in time is located at the bottom at a later point in time.

2.5.3 Under

2.5.3.1 HORIZONTAL RELATIONSHIP

The sentences in this group capture a non-canonical spatial configuration, in which the figure is located adjacent to the ground on a horizontal axis. It therefore departs from the vertical axis that is understood to characterise the underlying meaning of most spatial uses – and indeed some non-spatial uses – of *under*. This distinction is illustrated in Figure 10 and Figure 11.

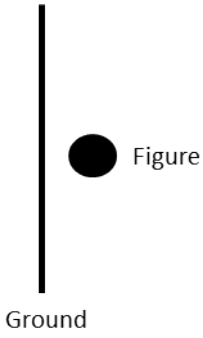


Figure 10 Horizontal relationship between figure and ground

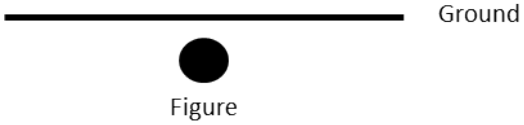


Figure 11 Canonical vertical relationship between figure and ground

While the configuration described by the sentences is a 90 degree rotation of the canonical *under* configuration, it does not lose the functional consequences associated with the canonical orientation, namely covering. The example *Rinse the dish under running water* perhaps exemplifies this most weakly, with the ground (*water*) unlikely to be covering the ground (*the dish*) in its entirety. Further, other speakers may challenge my judgment that *I'm a bit hot under the collar* and *They got under cover of the walls of the fortresses* should be categorised into this group, due to the idiomatic nature of the first sentence, and the collocation *under cover* in the second. These issues aside, the spatial configuration that underlies both sentences is the same as that which underlies the other members of the group.

2.5.3.2 ACCORDING TO

This set of non-spatial uses of *under* collectively capture a sense in which an action is permitted by a rule. The rule therefore exercises control over a particular action, in a way that we might liken to the control that a figure located immediately underneath a ground in a spatial configuration is subject to.

2.5.3.3 SUBJECT TO

This set of sentences each describes the exertion of force by an abstract ground on an abstract or physical figure. Just as I observed a relationship between the ACCORDING TO sense with the functional consequence of control inherent in the vertical configuration canonically associated with *under*, there is a relationship with that configuration here also. In this case, an abstract extension of the functional consequence of force exerted by a ground onto a figure is evident in these sentences. The force described by these sentences, for example in *Remember their hospital is under threat*, is typically negative. That said, I do not find that the distinction between negative and neutral forces, such as that exemplified by the sentence *The question of intercommunion is under discussion*, is sufficiently meaningful to license assigning this stimulus to a separate group.

2.5.3.4 VERTICAL RELATIONSHIP, NO CONTACT

This is the first of two groups in the stimuli that capture what is considered to be the canonical configuration that underpins *under*: a vertical relationship between a figure and ground, in which the figure is located in a position inferior to that of the ground. In this case, the group is further specified to capture only those examples in which the figure and ground are not in contact.

2.5.3.5 VERTICAL RELATIONSHIP, WITH CONTACT

These sentences capture a similar spatial configuration to those described by the sentences in the previous group, but are distinguished from those in that group due to the presence of contact between figure and ground. The sentences therefore not only describe a vertical configuration, but one which is also layer-like.

2.5.3.6 UNDER THE CONTROL OR AUTHORITY OF

The final non-spatial group in the stimuli captures a sense of *under* that describes the exertion of control by a human, or group thereof. In each case, the figure is positioned on an abstract vertical schema in a position inferior to the ground. The group is similar in meaning to the sentences in the ACCORDING TO and SUBJECT TO groups, which also capture an abstract extension of the functional consequence of the spatial configuration canonically associated with *under*.

2.5.3.7 A note on ACCORDING TO, SUBJECT TO and UNDER THE CONTROL/AUTHORITY OF

While I argue that these three non-spatial groups should be distinguished from one another, they do have something in common. Specifically, each of the three groups is characterised by abstract control or force exerted by one entity on another. For example, *You can emigrate to Britain under limited criteria* (ACCORDING TO) describes the force exerted by a set of rules on an individual. *We're under real pressure at the moment* (SUBJECT TO) describes an abstract force being applied to a figure. Finally, *We'll be taxed under the Conservatives* (UNDER THE CONTROL OR AUTHORITY OF) describes an impending force to be exercised upon a figure by a named individual or collective animate ground. Further, the ground can be understood to occupy a particular position in a type of hierarchy; specifically, one that is superior to that of the figure.

2.5.4 Above

2.5.4.1 BETTER THAN

These sentences capture a qualitative sense of *above*, divided into sentences in which one person measures something against a qualitative scale, or a person measures themselves against such a scale. I believe that it is based on spatial configuration canonically associated with *above*, in which a figure is located in a higher position than a ground. Further, I argue that it is an elaboration of the MORE THAN sense discussed below that is extended to describe qualitative scales.

2.5.4.2 MORE THAN

The sentences in this group describe the position on a quantitative scale; specifically, figures which are of higher value than a ground value. Like the sentences in the BETTER THAN group, I judge that this sense is semantically related to the spatial configuration typically associated with *above*. In their study of *over*, Tyler and Evans (2001) noted a correlation between vertical scales and quantity. This can be exemplified by considering what happens when a £1 coin is stacked upon another, and then another is added, then another, and so on. With each addition, the number of coins increases in correlation with the height of the stack. The same correlation is true of *above*, too. Due to the similarity of the spatial configurations that these words are typically associated with, both words can be used to describe numerical scales.

2.5.4.3 VERTICAL RELATIONSHIP

This set of sentences represents the spatial relationship canonically associated with *above*, in which a figure is positioned superior to a ground, and in which there is a vertical relationship between the two objects. The group does not specify for degree of distance between figure and ground, and includes sentences that describe close proximity, as in *There was a faint bruise above her eyebrow*, and those which describe great distance, as in *The stars above were partly obscured by clouds.*

2.5.4.4 TEXT USES

While the other five groups of stimuli in this task can be easily judged as representing either spatial or non-spatial uses of *above*, the sentences in the group I will label TEXT USES are trickier to judge. They elude single categorisation into one or the other domain: *The comments (listed above) are worrying* describes the position of comments both in space, in that they direct the reader to a location in a document, but also in time, in that they refer to a point in the past, at which the reader initially encountered them. For this reason, and the fact that they are typically encountered in a specific medium – written English – I anticipate that participants will agree that the sentences in this group are highly distinct from other stimuli.

2.5.4.5 VANTAGE

I judge that this set of examples describe an elaboration of the canonical spatial relationship underpinning *above*; while the figure is located in a position superior to the ground, the relationship is on a diagonal, rather than vertical, axis. In each case, the ground is an extended, three-dimensional space of which the figure has a vantage

perspective. This configuration is illustrated in Figure 12, in which the grey dashed lined indicate the range of perspective held by figure over ground.

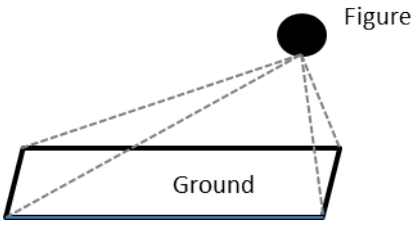


Figure 12 Schematic representation of VANTAGE.

2.5.4.6 HIERARCHY

Each sentence in this group describes the position of one person or post in a hierarchy in relation to another. In this case, the ground is positioned higher than the figure on an abstract vertical schema.

2.5.5 Below

2.5.5.1 WORSE THAN

This group of sentences describe opposite positions to those described in the BETTER THAN group in the *above* task. In this case, an entity – concrete or otherwise – is compared against a qualitative scale, and is found to be in an inferior position to a ground on that scale.

2.5.5.2 VANTAGE

While the sentences in the VANTAGE group in the *above* task described the superior position of a figure relative to a group, albeit on a diagonal axis, the opposite is true of the sentences in the VANTAGE group for *below*. In this case, the figure is located lower than the ground, as illustrated in Figure 13. Figures specified in these examples occupy extended three-dimensional spaces.

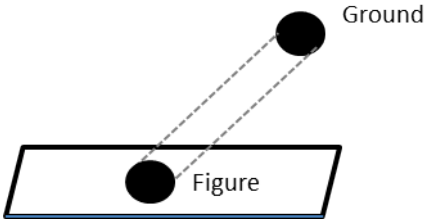


Figure 13 Schematic representation of VANTAGE. Dashed lines represent perspective of ground held by figure

2.5.5.3 UNDERNEATH (3D)

The sentences in this group describe a three-dimensional configuration in which a ground occupies an extended space underneath the figure. I make the distinction between the three spatial groups in the orientation of the figure and ground in each sentence. In this category, the figure and ground are oriented along a vertical axis, whereas sentences in the VANTAGE category describe a diagonal relationship. This relationship is illustrated in Figure 14.

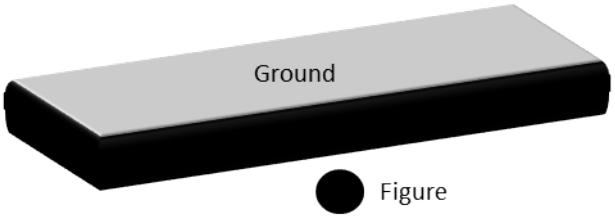


Figure 14 Schematic representation of UNDERNEATH

2.5.5.4 LOWER THAN (2D)

The members of this group contrast with those of the two other spatial groups, VANTAGE and UNDERNEATH in their dimensions: unlike those groups, the sentences in this group describe two-dimensional relationships between figures and grounds. This configuration is diagrammed in Figure 15.

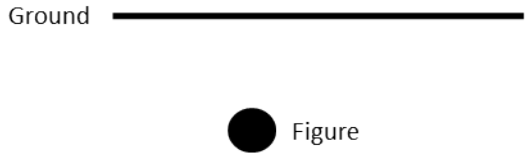


Figure 15 Schematic representation of LOWER THAN

2.5.5.5 LESS THAN

The members of this group are direct opposites of the members of the MORE THAN group in the *above* stimuli. Like that group, the sentences in the LESS THAN group describe a position on a quantitative scale in which the figure is of a lesser value than a ground value. I judge that, just as the MORE THAN sentences are related to the BETTER THAN sentences, the overarching meaning captured by this group is related to – but distinct from – that of the WORSE THAN group. This group does not require that the exact value of the ground is specified, but only that some position on a quantitative scale is identified. It is coincidental that all of the stimuli in this group describe positions on specifically financial scales, and I would judge that sentences that describe a lesser position on a more generic quantitative scale, such as *His share*

of the popular vote fell to below 50% in the provincial election of 1995 would be an equally good member of the group.

2.5.5.6 TEXT USES

We can compare the members of this group with those of the TEXT USES group in the *above* stimuli. In this case, the sentences refer to the position of figure on a page. While the examples in this group typically describe the position of text, the example *Fill in below all the tasks that you do in a typical day* seems an equally good match for the group, describing the intended position of some writing. As noted in my description of the TEXT USE group in section 2.5.4.4, this group defies easy categorisation as describing either a spatial configuration or metaphorical arrangement of figure and ground. They describe the lower spatial position of a figure relative to a ground, as well as the relative temporal relationship between figure and ground. Assuming that the document is read in a canonical, linear fashion, the figure text will be encountered at some point in the future. As noted earlier, the opposite is true of TEXT USES of *above*.

2.6 Data collection

In this section, I outline the approach I took to gathering data to answer the question of whether the senses that I find meaningful are meaningful in the minds of other speakers. In brief, it firstly introduces the methodology used. It then provides relevant information about the participants who completed the tasks. Thirdly, details of the stimuli used are provided, followed by detailed information about the task procedure. Finally, it introduces the statistical model used to measure the degree to which individual participants and I agreed about how the stimuli should be sorted.

2.6.1 Methodology: Sentence-sorting tasks

Sentence-sorting tasks offer the chance to observe the categories people make when they are asked to sort stimulus sentences according to a particular criterion. In the case of this research, participants are asked to sort them on the basis of the meaning of the capitalised target word, either *over, under, above* or *below*. In this study it is assumed that the categories people make are indicators of (some of) the senses they find meaningful. The task is straightforward, easy to understand, and participants report they find the tasks interesting and stimulating. The task is also very easy to publish online, which allows large and diverse samples of participants to be recruited. The results generated are relatively straightforward to make sense of:

between them, agreement statistics and data visualisation tools allow for powerful insights into how well individuals agree with each other, whether there are senses that are common to many speakers, and, if there are, whether or not these senses are related. Sentence-sorting tasks can be administered in two ways: participants may be given a set of sentences and asked to categorise them into groups of their own making. This is known as an open-sort task. A closed-sort task, in contrast, sees participants categorising sentences into predetermined categories. The closed-sort task method is adopted in this study and is described in section 2.6.1.1, below.

Sentence-sorting tasks have an established place in the study of word meaning. In work led by Rice to understand the polysemy of the words *at, on* and *in*, she and her colleagues carried out several open sort tasks. In each case, every participant sorted examples of each word (20 examples for each word in Sandra and Rice, 1995; and 50 each in Rice, 1996, and Rice, Sandra, and Vanrespaille, 1999). Divjak and Gries (2008) used a smaller task in their study of near-synonymous Russian verbs; in that case, each participant sorted nine sentences, representing examples of different near-synonymous verbs. The participants were initially asked to sort them into as many groups as they wished; in a second experiment, the same participants sorted the same sentences into up to three groups; and in the final experiment, the same sentences were sorted by the same participants into three groups, in which each group must comprise three different verbs. Hong and Baker (2011), in a WSD study, carried out a closed sort task comprising 18 sentences using one of two senses of *justify*. Participants were asked to review each sentence and allocate them to one of the two sense categories provided. In a very large task, Baker (1999) recruited participants to sort up to 244 examples of *see* in an open sort task. His closed sort task saw participants sorting 99 sentences into either 19 (experiment 1) or seven (experiment 2) sense groups.

Sentence sorting tasks have a clear overlap with the methodology frequently adopted in human word sense disambiguation (WSD) research. In those studies, trained or naïve annotators are presented with one example of a given word at a time, and asked to judge which of a set of predetermined sense tags best matches the sense of the target word. In some cases, annotators are permitted to use multiple tags (e.g., Véronis, 1998). Clearly there is some difference in the delivery of these tasks, in that

the tasks used here present every stimulus sentence at once, but the underlying principle remains the same.

2.6.1.1 Closed-sort tasks

In a closed-sort task, participants are presented with a set of stimuli and a number of preset categories. They are asked to sort the stimuli into the categories according to a given rule. In this research, participants are given a set of sentences, each of which uses a target word in a particular sense. The categories they are asked to sort the sentences into are labelled with a sentence that also features the target word being used in a particular sense. Each category represents a different sense of the same target word. Participants are asked to judge what the capitalised target word means in each sentence, and sort it into a category with the corresponding meaning. The intended final result is that each category should consist of a set of sentences in which the sense of the target word is the same, and which is the same as the sense of the target word in the category label.

The purpose of this type of closed-sort task is to understand how similar participants and I are in our judgments of which sentence belongs in each group: it allows me to measure how well each participant and I agree that a particular sentence should be assigned to a particular category. It is, therefore, a good means of testing how well my intuitions about word senses match those of others. A weakness of the task, however, is that because the task is very structured, and because the category label sentences provide some guidance about which sentences each category should contain and are contrasted against labels of other categories, participants may sort the sentences in a way that indicates high agreement with my word senses. Without these guides, they may sort them differently. There are two obvious ways of overcoming this weakness. First, distractor sense categories can be provided. They would be labelled with sentences which use a sense of the target word judged by the experimenter to be absent in the set of stimulus sentences. Alternatively, participants could be given a set of sense categories without distractors, but informed that they may use as many or as few as they wish. The latter approach is taken in this research.

The categories given to participants reflected what I considered to be six distinct senses of each of the target words represented in the stimulus sentences. Each sense,

in my view, was used in six stimulus sentences. The six sentences corresponding to each category will be henceforth called *target sentences*, and the category to which I believe they belong will be referred to as the *target category*. Accordingly, each participant was asked to sort a total of 36 sentences – each examples of one of *over, above, under* or *below* – into at least one of the six predetermined groups. The task is consequently highly structured: stimuli have been selected specifically to represent what I judge to be a particular sense, and there are an equal number of exemplars of each sense. The stimuli and categories are shown in Table 1 to Table 4.

2.6.2 Participants

A total of 298 native English-speaking participants completed the tasks. Participants were primarily recruited online using the Reddit Sample Size website. Sample Size is a forum in which researchers can recruit participants for experiments, surveys and other research projects. A small number of participants were recruited by emailing students in the Faculty of Arts, Design and Social Sciences at Northumbria University, and through personal contacts. Participants were not rewarded for taking part.

Given that the purpose of this study is to understand the extent to which the sense distinctions that I – as a linguist – find meaningful correspond to those meaningful in the minds of other speakers, I was also interested in understanding how similar they were to the intuitions of other linguists. For this reason, I sought participants who had experience of studying linguistics. In this study, participants who have either completed a PhD in linguistics or are currently working on a PhD in linguistics are classified here as linguists. Information about the participants is summarised in Table 5.

Table 5 Experiment 1 Participants

	Total number participants	Non-linguists	Linguists	UK participants	Non-UK participants
Over	79	76	3	14	65
Under	62	59	3	14	48
Above	91	88	3	10	81
Below	66	60	6	13	53

Participants had a range of educational backgrounds. All had completed or were in the process of completing at least a high school qualification, and the highest level of

education was a doctoral degree. Non-student participants represented a wide array of job types.

2.6.3 Stimuli

Stimuli consisted of 36 examples of one of *over, under, above* or *below*. Sentences were extracted from the internet and from the spoken and written sections of the British National Corpus, edited to make the examples well-formed sentences. Where appropriate, the sentences were edited to reduce their length. This decision was made to ensure that as many sentences were visible on the computer screen at a time.

The sense categories were labelled with a sentence, also an edited corpus or internet extraction, exemplifying each particular sense.

2.6.4 Procedure

Participants recruited online completed the sorting task using an online virtual card sorting tool called OptimalSort. Participants recruited in person completed the task using a set of cards, and some were also asked to describe the meaning captured by each group. Due to time constraints it was not possible to elicit this further data from all participants who completed the task in person. The procedures for these two test conditions are described fully in sections 2.6.4.1 and 2.6.4.2.

2.6.4.1 Online sorting task

Prior to starting the task, information about the task and how participants' responses would be stored was presented on the screen. Participants were required to read this and confirm that they understood it, and provide their informed consent to participate in the research. After consent had been received, participants were shown written instructions about how to complete the task (see appendix 1 for a copy of these instructions). Briefly, participants were instructed to sort a set of sentences into one or more of a set of predetermined groups. They were explicitly instructed to sort sentences on the basis of what the target word – which was capitalised – meant. It was made clear that the goal of the task was to sort all of the sentences into groups in which the meaning of the capitalised target word was the same in each member of the group. The instructions disappeared when the participant moved the first sentence, but could be recalled at any time.

After being presented with the task instructions, the task was revealed on the screen. Stimulus sentences were presented in a column on the left side of the screen, with the predetermined categories positioned in a larger sorting pane to the centre and right of the screen, as shown in Figure 16.

Figure 16 Screenshot showing online sorting task before any sentences have been sorted

Participants were advised to read through all of the stimulus sentences and the sentences used to label the six categories, considering carefully what the target word, which was shown in full capitals, meant in each case. Afterwards, they should move each sentence into a category which used the same sense of the word. Sentences were moved by dragging each one and dropping it onto the category label in the sorting pane, as shown in Figure 17, below. Sentences could be moved into a different category by dragging and dropping it into a new category.

Figure 17 Screenshot showing online sorting task after some sentences have been sorted

Sentences could be moved in and out of categories until the participant was satisfied with their sorting decisions. Once they were satisfied with their groups, they clicked a button to indicate that they had finished the tasks. The results were then stored and accessible in the back end of the programme. Participants were required to sort all of the stimuli before they could submit their responses.

2.6.4.2 Face-to-face sorting task

Printed information about the task which matched that given to participants who completed the task online, was given to participants prior to starting the task. Their informed consent was collected, and then participants were given printed instructions. The instructions concerning the sorting task were identical to those provided to participants who completed the task online; however, a subset of seven participants received instructions that also included an instruction to describe the meaning of each group at the end of the task.

While participants read the instructions, six sheets of white A4 paper, at the top of which were printed one of each of the six category label sentences, were laid out in a random order in front of a participant. A set of cards, on which the stimulus sentences were printed, were shuffled and set in a stack face down in front of the participant. The participant was asked to read the instructions in full and ask any questions they wished. Participants were able to ask questions while completing the task, but no guidance as to how the cards could or should be sorted was given.

Participants were instructed to place each printed card onto one of the six A4 sheets. They were able to move the cards in and out of different groups until they confirmed that they were happy with their decisions. A subset of seven participants were then asked to tell me what they considered to be the underlying meaning of each group. Their responses were typed as close to verbatim as possible.

Once the participant had completed the task, their responses were recorded on an Excel spread sheet and later input into the OptimalSort programme so they could be included with data gathered online.

2.6.5 Statistical analysis

The present study uses Cohen's kappa to measure agreement between each participant and me. The statistic calculates how well individual participants agreed with the sentence categorisations I make as shown in Table 1 to 4. Cohen's kappa is a well-established means of calculating pairwise agreement of nominal categorisation, returning values from -1.0, representing total disagreement, through 0, representing agreement that would be expected by chance, to 1.0, representing perfect agreement. Cohen's kappa shows the *magnitude* of agreement, and does not return significance values.

Cohen's kappa is considered to be the "most widely accepted measure of inter-rater reliability … especially in the medical literature" (Sun, 2011, p. 146), and has particular use in diagnosis. A meta-analysis of the use of kappa in the classification of pressure ulcers in 15 studies showed that kappa values in these studies ranged from 0.15 to 0.97 (p. 152-3), indicating "significant heterogeneity across studies" (p. 156). It is a versatile measure, though, and has been used in other fields including the study of the strength of differentiation of well-formed and not well-formed usages of an artificial grammar by marmosets and macaques (Wilson et al., 2013); this study returned values of 0.67 for the coding of degree of differentiation by macaques, and 0.39 for the degree of differentiation by marmosets. A study of iris colour judgments based on photographs by Seddon et al. (1990, p. 1597) returned an agreement value of 0.76.

What constitutes an acceptable level of agreement is the subject of disagreement – an irony that appears to be lost on those who posit acceptability ranges. Interpretation of kappa values is typically based on the scales proposed by Landis and Koch (1977, p. 165), who propose that a kappa value of 0.41-0.6 reflects moderate agreement, 0.61-0.8 reflects substantial agreement, and 0.81-.1.0 reflects almost perfect agreement. However, the authors acknowledge the arbitrary nature of these classifications. In computational linguistics literature Artstein and Poesio (2008) have described the interpretation of agreement scores as "little more than a black art" (p. 576). They note that agreement scores in computational linguistics research tend to follow the interpretation conventions adopted in content analysis, in which values of 0.8 or higher constitute good agreement, and in which tentative conclusions may be drawn

from values between 0.67 and 0.8. However, they observe that other authors propose more stringent interpretations; Neuendorf (2002, p. 3) recommends considering values of 0.9 or more as acceptable all in situations, 0.8 to 0.9 as acceptable in most situations, and scores less than 0.8 constituting great disagreement.

The importance of agreement is clearly relative; it is important that a pair of diagnosticians charged with assessing what investigations a patient needs based on initial presentation at hospital reach a very high level of agreement, thus necessitating that the lower bound of acceptable agreement is set high. In situations where poor agreement would have less serious consequences, the boundaries of what constitutes an acceptable level of agreement are perhaps more flexible. Artstein and Poesio (2008) recommend that rather than a single cut-off point is used, agreement values are considered for their acceptability in terms of the goal of the task. They describe, for example, an annotation study in which the quality of annotation was only useful where agreement was in excess of 0.8.

It is not the goal of this project to identify the senses of *over, under, above* and *below* for practical application, but is instead interested in studying inter-speaker variation for its own sake. For that reason, identifying an acceptable level of agreement is not completely necessary. Of course, I am interested in establishing whether individual participants and I have very high agreement, or very poor agreement, and it is therefore useful to have a scale that captures what constitutes poor, good and excellent agreement is. However, the availability of agreement scores in and of themselves offer a scale that I can interpret; for example, comparing two pairs of participants' scores, one of 0.8 and one of 0.4, I will be able to say with confidence that the first participant and I have more similar intuitions about the senses of the relevant target word than the second participant and I do.

In summary, Cohen's kappa is used for categorical data collected from two independent participants. It has been shown to return a broad range of values in research across disciplines, and a tentative interpretation of values is that a kappa of approximately 0.8 or greater is acceptable.

2.7 Results and discussion

This section presents statistical and qualitative analyses performed on the sentence-sorting data. It begins by presenting agreement values, showing how well participants and I agreed with each other. It then addresses the possibility that the two experimental settings used – face-to-face and online – might have affected how participants "performed" in the task, which might be reflected in the degree to which we agreed with each other. It then moves to a more qualitative analysis, and discusses popular placement matrices produced for each task.

The kappa values returned, the averages of which are at least close to the "acceptable" value of 0.8 (Neuendorf 2002), as well as the popular placement matrices shown later in this chapter, indicate that participants sorted the sentences in systematic ways. This indicates these participants are sensitive to meaningful distinctions in the way the words *over, above, under* and *below* are used. This complements findings by Rice and her colleagues suggesting that participants can categorise examples of polysemous words based on their meaning (Cuyckens et al., 1997; Rice et al., 1999; Sandra and Rice, 1995).

In the following sections, I will discuss observations that are of most relevance to the aim of this chapter. I will open by presenting and interpreting kappa values showing how well individual participants and I agreed about how examples of these four words should be categorised. I will then consider in detail the ways in which examples of each of the four words were sorted, paying particular attention to convergence and divergence between my own sense distinctions and those of the participants. I will then discuss observations made across all four tasks about differences in lumping and splitting tendencies. I will close by discussing the contribution the study makes to our knowledge about and the debate over the role of linguists' intuitions.

2.7.1 How well do participants and I agree about how the sentences should be sorted?

Participants' responses were coded to specify which category each sentence had been sorted into. Individual participants' data, each individually paired for comparison with my sense distinctions, were run through R using a script featuring

the cohen.kappa function in the psych package (Revelle 2015). The kappa values returned are summarised in Tables 6, 7, and 8.

Table 6 Summary statistics of pairwise Cohen's kappas for all participants, to 2 significant figures.

	Mean	SD	Range	Min.	Max.
Over	0.88	0.08	0.40	0.60	1.0
Under	0.72	0.15	0.73	0.20	0.93
Above	0.76	0.10	0.50	0.47	0.97
Below	0.75	0.11	0.50	0.40	0.90

Table 7 Summary statistics of pairwise Cohen's kappas for non-linguist participants, to 2 significant figures.

	Mean	SD	Range	Min.	Max.
Over	0.88	0.08	0.40	0.60	1.0
Under	0.72	0.15	0.73	0.20	0.93
Above	0.76	0.10	0.50	0.47	0.97
Below	0.75	0.11	0.50	0.40	0.90

Table 8 Summary statistics of pairwise Cohen's kappas for linguist participants, to 2 significant figures.

	Mean	SD	Range	Min.	Max.
Over	0.87	0.04	0.07	0.83	0.90
Under	0.69	0.21	0.40	0.53	0.93
Above	0.72	0.17	0.34	0.53	0.87
Below	0.78	0.09	0.20	0.67	0.87

Three patterns of variation in agreement values are observed in these tables: variation in agreement across the four words, between the four words, and across participants. Each will be addressed in turn, and will focus specifically on the results presented in Table 6, which capture agreement values for all participants.

First, there is variation across the four words. If we compare the mean values in the grey-shaded column for each word, we can see variation in average agreement values across the four tasks. While examples *over* tend to have been sorted in ways that are broadly similar to the way I would classify them, as reflected by a mean agreement value of 0.88, examples of *under* are, on the whole, sorted rather differently, as reflected by a mean agreement value of 0.72.

Next, there is variation in agreement within words. If we consider the 'Max.' column and *over* row in Table 6, we observe an agreement value of 1.0, demonstrating that at least one participant who completed the *over* task reached complete agreement with me about how the stimuli should be sorted. In contrast, if we consider the 'Min.' column for the same word, we can see that one participant and I achieved an

agreement score of 0.6, indicating some disagreement about how we should categorise the sentences. The same pattern is seen in the rest of the table: there are instances of both high and low agreement values generated in all four tasks.

Finally, by considering the values 'Range', 'Min.' and 'Max.' columns in Table 6, Table 7 and Table 8, we can observe variation across participants; i.e., there is variation in the extent to which individual participants and I agree about how the stimuli should be sorted. For example, if we look at the cell highlighted in red in Table 6, we can see that one participant and I achieved an agreement score of 0.20, demonstrating that we sorted the stimuli very differently. In contrast, the cell highlighted in green shows that a participant in the same task and I agreed with each other very well, achieving an agreement score of 0.93. Likewise, the 'Range' column shows how much variation there was in my agreement with all participants.

The fact that perfect agreement was rare, and that average agreement was typically in the mid 0.7s, suggests that there are differences in the way participants and I categorise examples of these words, therefore indicating that participants and I differ in which sense we believe each sentence to exemplify. Put simply, the lack of robust agreement in the data suggest that the senses presented to participants, i.e., those which I find to be meaningful, do not always align with those of other native speakers of English.

Let's now move to compare how well linguist participants and I agreed with my agreement with non-linguists. As shown in the 'Mean' columns in Tables 7 (non-linguists) and 8 (linguists), with the exception of the *below* task, on the whole I reached a higher level of agreement with non-linguists than I did with linguists. This is a somewhat surprising outcome; if we were to expect any group of participants to agree with an expert's intuitions, it would be a group of experts, rather than a group of laypeople.

2.7.1.1 Did using two different experimental settings affect participants' performance?

While the fundamental procedure used in both test conditions was the same, using two different approaches to data collection introduces the possibility that participants in the two conditions performed the tasks in different ways. Participants who

completed the task in front of me may have been inclined to take the task more seriously than those who completed the task online, as the face-to-face condition has a more performative quality. In contrast, those who completed the task online were not under the same pressure to "perform" in this way. Moreover, since some participants in the face-to-face setting were told that they would be asked to describe the meaning of each group, they may have considered their sorting decisions more carefully than those who completed the task online, as they were aware that they would be required to defend their categorisations. While using two different test settings does introduce the possibility of a confounding variable, I judged that the benefit of using two experimental settings outweighed the risk of a confound. Specifically, expanding the data collection to recruit participants online allowed me to recruit a very large number of participants representing a diverse demographic, which would not have been possible had I recruited all participants to complete the task in person. Using a face-to-face approach also opened up the opportunity for studying participants' thought processes used when completing the task. Accordingly, the think aloud protocol was used to record verbalisations of the decision processes participants used, and to record whether they and I had similar intuitions about what each group meant. This data is not presented here as the content was highly predictable, and provided no significant additional insights into how well participants' sorting decisions converged with, or diverged from my own. When participants and I tended to agree on how a particular set of sentences should be sorted, we tended to describe the meaning in similar ways. Equally, when we disagreed over how some sentences should be sorted, we also disagreed about the meaning captured by that group.

To establish whether the use of two experimental setting conditions affected participants' agreement scores, the Wilcoxon-Mann Whitney test was used to compare Cohen's kappa scores across the two groups of participants. This test indicates that there are no significant differences in the extent to which participants in the two setting conditions agreed with me ($U(296) = 2386.5$, $W(296) = 2596.5$, $Z = -1.062$, $p = 0.288$).

2.7.1.2 The effect of task structure

In a sorting task, one would expect some degree of disagreement due to a range of factors including fatigue, error, and failure to use the sorting variable instructed. What we see here appears to go beyond that, however: there are cases in which agreement, while above chance, is really rather low. In the context of this study, the degree of such poor agreement is rather surprising. The task is on a relatively small-scale, and has a great degree of structure, as described in section 2.6.1. Firstly, participants must use pre-determined classification groups which collectively hint at six distinctions that I judge to be present in the examples of the target words being sorted. Secondly, each predetermined category reflects a sense which I judge to be used in six of the stimulus sentences. If a participant observed that a particularly clear sense distinction – the TEXT USE sense of *above* and *below*, for example – corresponded to six stimulus sentences, it would not be a mean feat to ponder whether, given that each task consists of 36 stimulus sentences and six categories, each category should end up with six members. This would be a feat that becomes much more straightforward if a participant found two sense categories to be very distinct and easily accommodate six stimulus sentences apiece. Despite the structured nature of the task, there remains some disagreement with my categorisation decisions. This indicates that although the structured nature of the task should have helped participants to produce a sorting solution compatible with mine, it does not appear to have done. This strongly suggests a mismatch between my intuitions and those of the participants.

2.7.2 How were examples of each word sorted?

2.7.2.1 How data were analysed

This section is based upon analysis of popular placement matrices generated by the OptimalSort programme. These matrices show the percentage frequency with which the participants as a group assigned each stimulus sentence to each category. The matrices are constructed using OptimalSort's proprietary algorithm and groups together sentences which are sorted into the same category with highest frequency. An annotated example is given in Figure 18. Groups of sentences that the algorithm has calculated to have been sorted together and into a particular category with highest frequency are highlighted in blue. The purple box shows sentences that have been sorted together with a high degree of agreement, with the red boxes in and around it highlighting the infrequency with which sentences in this category are

sorted into other groups. This is an example of what I shall describe as a *discrete* group. The yellow box shows sentences which have been sorted into groups with varying degrees of agreement across the participants. These groups show what I will describe as a high degree of *overlap*. Overlap is meant here to describe the fact that members of the strongest categories identified by the OptimalSort grouping algorithm (shaded in blue in the matrix) are judged by some participants to be members of other categories. This is shown in the percentage values outside of each blue box.

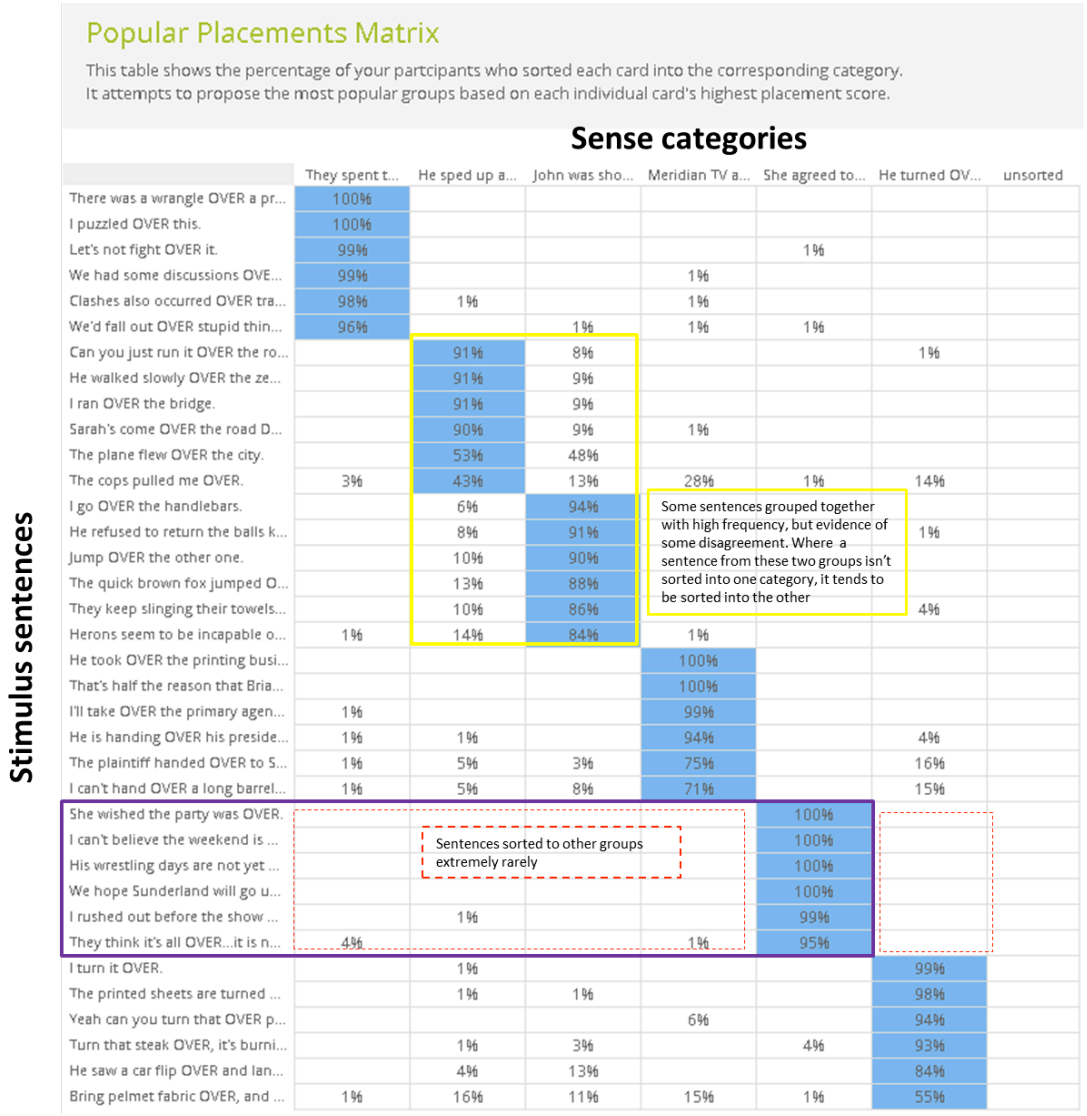


Figure 18 Annotated popular placement matrix for the *over* task.

Analysis of the popular placement matrix allows a quantitative approach by considering the frequency with which particular sentences were assigned to particular groups, as well as a qualitative approach by studying which sentences were allocated to each category and discussing why particular sorting decisions may

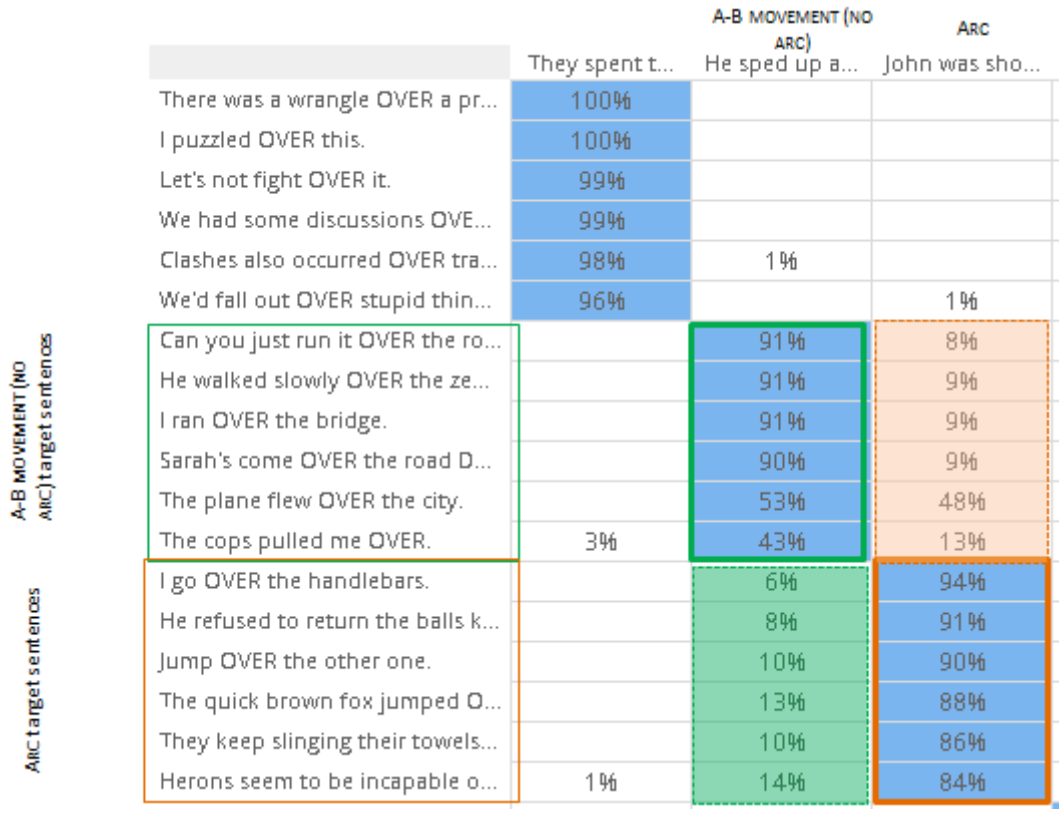
have been made. This analysis serves to inform and elucidate the kappa values presented in Table 6. Kappa values provide meaningful and useful information, and the following close analysis is intended to explain those values. Jointly, it is hoped that they will provide an informative examination of the decisions the participants made. The focus of this chapter and the following analysis is on a comparison of participants' sorting decisions and my intuitions about how the stimuli should be sorted. However, given that the matrices show the way sentences were sorted by participants as a group, the analysis will consequently also consider sorting tendencies across the participant groups.

Popular placement matrices are available in full in Appendix 2, and, where the written analysis may benefit from illustration, snapshots of relevant sections of the matrices are presented in the body of the text.

2.7.2.2 Over

As reflected in the relatively small range of agreement values shown in Table 6, the way in which examples of *over* were sorted is fairly consistent across individuals, manifesting as reasonably discrete groups in line with my sense distinctions.

There is, however, a limited degree overlap between two of the spatial groups – those describing arc-like trajectories, and those describing what I judge to be arc-less movement from one place to another. While this tendency is limited, it merits discussion. This overlap suggests that while participants tend to agree with my intuition that there is a salient distinction between these two trajectories, some do not. Where participants do not share this intuition, their sorting decisions systematically differ from mine. Specifically, when they do not categorise a target sentence to the target category, with the exception of *The cops pulled me over*, they almost always assign the target sentence to the other category. In other words, when the sentence *The plane flew over the city* isn't assigned to the target A-B MOVEMENT (NO ARC) category, it is typically assigned to the ARC category. This suggests that the ARC sense may be a special case of the sense that I have labelled A-B MOVEMENT (NO ARC). This is illustrated in the snapshot of the similarity matrix shown in Figure 19.



		A-B MOVEMENT (NO ARC)	ARC
	They spent t...	He sped up a...	John was sho...
There was a wrangle OVER a pr...	100%		
I puzzled OVER this.	100%		
Let's not fight OVER it.	99%		
We had some discussions OVE...	99%		
Clashes also occurred OVER tra...	98%	1%	
We'd fall out OVER stupid thin...	96%		1%
A-B MOVEMENT (NO ARC) target sentences: Can you just run it OVER the ro...		91%	8%
He walked slowly OVER the ze...		91%	9%
I ran OVER the bridge.		91%	9%
Sarah's come OVER the road D...		90%	9%
The plane flew OVER the city.		53%	48%
The cops pulled me OVER.	3%	43%	13%
ARC target sentences: I go OVER the handlebars.		6%	94%
He refused to return the balls k...		8%	91%
Jump OVER the other one.		10%	90%
The quick brown fox jumped O...		13%	88%
They keep slinging their towels...		10%	86%
Herons seem to be incapable o...	1%	14%	84%

Figure 19 Snapshot of popular placement matrix for *over* showing percentage frequencies with A-B MOVEMENT (NO ARC) and ARC sentences were sorted into their respective target categories

This overlap may be attributed to the fundamental similarity in these trajectories: both feature the movement of an object from one position to another, as illustrated below:

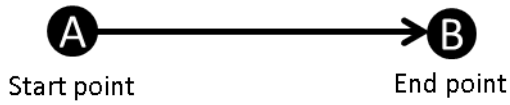


Figure 20 Motion configuration underlying the A-B MOVEMENT (NO ARC) sense of over

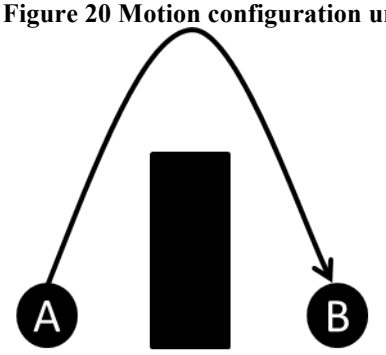


Figure 21 Motion configuration underlying the ARC sense of over

Figure 20 illustrates the linear movement from one point, A, to a second point, B. Figure 21 shares this start- and finish-point, but the shape of the trajectory is an

elaboration of that in Figure 20, taking into consideration an obstacle – in the target sentences, obstacles include handlebars and a fence. The target sentence *They keep slinging their towels over the bedroom door* is one which is perhaps least exemplary of the six sentences; while in the other sentences an object moves location from one place to another, passing over an obstacle in doing so, part of the objects named in this sentence – the towels – move only in part: one end of the towel will remain on one side of the obstacle – the door – while the other will move to the opposite side. This is illustrated in Figure 22. It is interesting to note that the final position of the figure in this sentence corresponds to the paths described by the other members of the group.

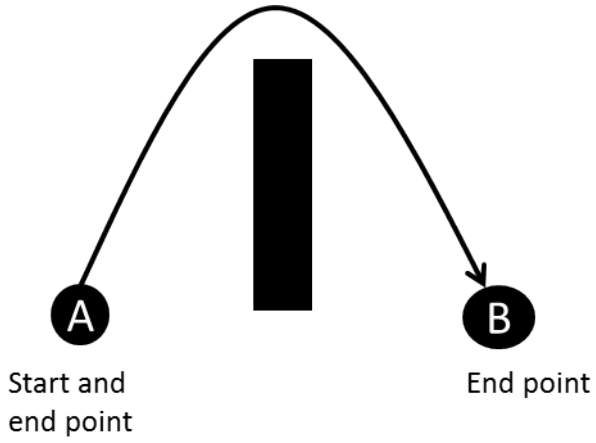


Figure 22 Motion configuration underlying the example *They keep slinging their towels over the bedroom door*

Evidence in support of the interpretation that the ARC sense is an special case of the sense that I have labelled A-B MOVEMENT (NO ARC) comes from the frequency with which the sentences *The quick brown fox jumped over the lazy dog* and *Herons seem to be incapable of stepping over the deterrent* are moved into the non-target A-B MOVEMENT (NO ARC) category, instead of the target ARC category (13% and 14% of participants, respectively). These sentences contrast with the other target sentences in that they describe movement over much flatter obstacles than those described in the other four sentences. It seems that, as the relative *flatness* of the obstacle increases, the tendency to sort the sentences into a category describing linear, horizontal movement increases also. Further complementary evidence in support of this interpretation is the greater tendency for participants to allocate the sentence *They keep slinging their towels over the bedroom door* to the target ARC category rather

than the non-target A-B MOVEMENT (NO ARC) category. In the case of the sentences describing the quick brown fox and the herons, the *movement* described is more similar to that which is described by the other ARC target sentences, such as *I go over the handlebars*, than the movement described in the towels sentence is to the other target ARC sentences. But while the sentences *Herons seem to be incapable of stepping over the deterrent* and *The quick brown fox jumped over the lazy dog* have more in common with the other target sentences in the movement they describe, the sentence *They keep slinging their towels over the bathroom door* refers to a ground and steeper arc that is more similar to the other target sentences; it is this which presumably licenses its membership in the ARC category. It seems, therefore, that the ARC category is a special case of the category that I have labelled A-B MOVEMENT (NO ARC): where movement is over a flatter obstacle, if it isn't assigned to the target category, it is assigned to the A-B MOVEMENT (NO ARC) category. Where movement is over a steeper obstacle, it is assigned to the ARC category. In the minds of the participants who adopted this sorting strategy, we might conclude that they have an overarching A-B MOVEMENT sense in which the arc of the movement is unspecified, and that the ARC sense is a sub-type of this sense.

Also worthy of attention is the range in agreement over where some target sentences from the TRANSFER category should be sorted. The popular placements matrix, a snapshot of which is shown in Figure 23, shows that the first four examples (*He took over the printing business, That's half the reason that Brian Tolbrook took over at Tettron, I'll take over the primary agenda*, and *He is handing over his presidency*) have very high levels of agreement that they belong in this category. This indicates that my intuitions about the meaning captured by these particular sentences are very similar to those of most of the participants. The final two examples, *The plaintiff handed over to Samuel Revill the first note,* and *I can't hand over a long barrelled weapon to that officer* have lower levels of agreement, and there is evidence that some participants believe they are more exemplary of the FLIP category.

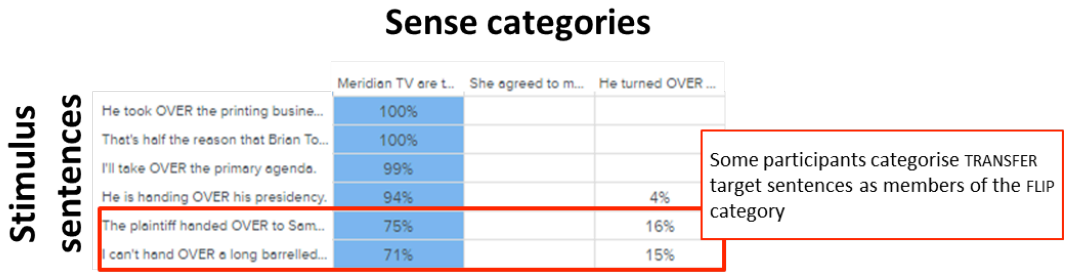


Figure 23 Snapshot of popular placement matrix for *over* showing categorisation of TRANSFER target sentences into the FLIP category

There is a division between the first four and last two examples in the objects that they describe as being transferred: while the first four examples describe the transfer of an abstract entity (a presidency, for example), the latter two describe the transfer of a physical object. While most participants agree that all six examples belong in the same group, and in a group that describes the transfer of an object that is underspecified in its concreteness, a small minority divide the group and allocate examples describing the transfer of a physical item to the FLIP category. Fifteen participants sorted at least one of these items into the FLIP category; twelve being American. This demographic information might explain this unexpected sorting pattern. A random sample of 100 examples from the British National Corpus (BNC) and Corpus of Contemporary American (COCA) reveals a marked distinction in the way a synonym of *flip*, namely *turn over*, is used. The BNC sample featured 11 examples of *turn over* used to describe a transfer, such as *The faster US troops can turn over responsibility for keeping the peace to Panamanian forces, officials said, the better*, versus 55 describing a flipping manoeuvre, as in *Now turn over the cards so that their backs are showing,* in addition to 33 examples of the phrase being used in other senses, for example *if you just would be so kind turn over to number three so I can have a look at some adverts*. In contrast, the COCA sample featured 58 examples used to describe transfer, as in *He was sent there to turn over the keys to the family car*, versus 20 to describe a flip, for example *As you cook, turn over the slices a few times*, and a further 21 used in other senses, such as *The neighbor down the street finally got his car to turn over.* In summary, while *turn over* is more frequently used to describe a flipping manoeuvre in British English, it is used more often to describe a transfer of a physical or abstract entity in American English. This finding suggests that in these particular cases, where participants and I do disagree

about how the sentences should be sorted, this may be the result of differences in the varieties of English participants and I speak.

In summary, there is some overlap across two of the spatial sense categories, reflecting disagreement – between individual participants and I, and between participants themselves – about which stimulus sentences constitute examples of each sense. There is interesting variation in the sorting of some non-spatial sentences that seems to be related to the variety of English the participant speaks. On the whole, there are discrete groups suggesting that the semantic boundaries represented by the preset sense categories are clear and distinct to other speakers. The matrix complements the generally high kappa values observed in Table 6. Collectively these analyses encourage the interpretation that participants' and my intuitions about the senses of *over* captured by these sentences are well aligned.

2.7.2.3 Under

The popular placements matrix on the following page shows that there is a large degree of overlap, with sentences being allocated to target- and non-target categories by large numbers of participants. These values show that while a number of relatively strong groups do exist, there are some sentences that a good proportion of participants think belong in another group. Take, for example, the sentence *He felt warm under the blanket*. This was sorted with highest frequency into the VERTICAL RELATIONSHIP, WITH CONTACT group, but 32% of participants assigned it to the HORIZONTAL RELATIONSHIP category, and 13% to the VERTICAL RELATIONSHIP, WITHOUT CONTACT category. This suggests that there is a large degree of uncertainty over which category some stimuli should be sorted into. This outcome corresponds to the generally rather poor agreement values shown in Table 6, and the wide range of agreement values. This outcome suggests that participants not only disagree with me on where some sentences should be sorted, and therefore what sense of *under* they use, but they also disagree with each other.

Figure 24 Popular placement matrix for *under*.

The popular placement matrix reveals a large amount of overlap between the groups describing spatial configurations. In particular, as shown in the box marked 1, target sentences for the VERTICAL RELATIONSHIP, NO CONTACT (hereafter NO CONTACT) category are frequently sorted into the VERTICAL RELATIONSHIP, WITH CONTACT (hereafter, CONTACT) category. Some participants sort target sentences from the CONTACT category into the NO CONTACT category, but this is less frequent. This suggests that there is disagreement between me and individual participants, and between the participants as a whole, over whether the distinction between vertical relationships with and without contact is a meaningful one.

Perhaps more interestingly, the higher frequency with which NO CONTACT target sentences are sorted into the CONTACT category suggests that vertical relationships

with contact between a figure and ground may be more basic than those without contact. This further suggests that a sense of *under* describing a vertical relationship without contact is simply an elaboration of one *with* contact – and one that speakers do not always find necessary to disambiguate (cf. Ide and Wilks 2007, p. 66, who discuss the necessity of fine-grained senses in human word sense disambiguation). We can compare this conclusion with Evans and Tyler's (2005) diagram of what they propose to be the protoscene underlying the basic spatial meaning of *under*, replicated here in Figure 25:

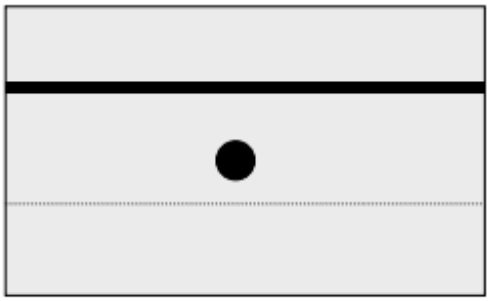


Figure 25 Protoscene for *under* (Evans and Tyler, 2005, p. 37)

Based on their proposal, it is reasonable to predict that the basic spatial sense of *under* is characterised by a vertical configuration in which figure and ground are not in contact. These findings challenge that prediction.

There is evidence of further uncertainty around which category the example *They got under cover of the walls of the fortresses* fits into. A small number of participants (13%) allocated it to the target category HORIZONTAL RELATIONSHIP, but most (43%) allocated it to a non-spatial category, specifically SUBJECT TO. It seems that while this example describes a spatial configuration equivalent to those described in the other sentences in the target category, namely, a horizontal covering relationship, the functional consequences of this particular scene – protection by virtue of being covered – appear to be judged to be of more importance, manifesting as being sorted most often with examples of figures being subject to the force of something else. In this case, the wall exerts a protecting force. However, the fact that there is a reasonably large group of participants (38%) who sorted the sentence into other non-target, but nonetheless spatial categories indicates that spatial nature of the sentence is a salient feature. This indicates that disagreement between participants as a group,

and between individual participants and I might be explained on the basis that we attend to different features of the configuration when judging what *under* means in this particular sentence.

The matrix as shown in Figure 24 shows that my and participants' intuitions diverge over where sentences should be sorted, regardless of whether the stimuli describes a spatial configuration or uses a non-spatial sense. Uses of *under* to describe non-spatial relationships are also subject to disagreement, with almost every non-spatial sentence being sorted into a non-target category by at least one participant.

To summarise, we have observed divergence in the way participants and I think that examples of *under* should be sorted. This suggests that there is divergence in participants' and my intuitions about what *under* means in each sentence. This disagreement is not limited to particular meaning domains: there is disagreement amongst both spatial and non-spatial sense categories. This observation is consistent with the generally low kappa values presented in Table 6. There is, however, evidence of some agreement with my senses. This is evident in the popular placements matrix, which shows that almost all of the sentences tended to be sorted into the target categories by at least half of the participants. There are, though, some exceptions, in which sentences are allocated to a non-target category more often than they are to the target category. For example, *They deny they were under any duty to offer advice* is allocated to the target category by only 37% of participants.

It is unclear what is behind the lack of consensus over where stimuli should be sorted. It may be explained on the basis that participants did not find the preset groups available to be particularly good matches for all of the stimuli, suggesting incompatibility between my intuitions about the senses used in these examples and those of other speakers. The possibility that the stimuli were not judged to match particularly well with a particular group further indicates that some examples are non-exemplary exemplars of particular categories. Returning to the sentence *They deny they were under any duty to offer advice*, it is not the case that participants simply allocated this stimulus to a particular non-target category. Instead, it is allocated to one non-target category (ACCORDING TO) by 40% of the participants, and to another non-target category (UNDER THE AUTHORITY/AUTHORITY OF) by 24% of

participants. The lack of decisiveness amongst this group of participants suggests that no one of these preset groups is a particularly good match. It may be the case that my intuitions about what sense this example represents simply doesn't match that of other speakers, who collectively share the same intuition about what sense it exemplifies.

2.7.2.4 Above

High values in the popular placement matrix show that there are pockets of good agreement with my sense distinctions across the participants. In particular, the TEXT USES of *above* constitute a very discrete group in near total agreement with my intuitions, with allocation of non-target sentences into the category, and allocation of target sentences to other categories, being an exceptionally rare occurrence. This is shown in the snapshot of the popular placement matrix in Figure 26. The same is true of the HIERARCHY target sentences.

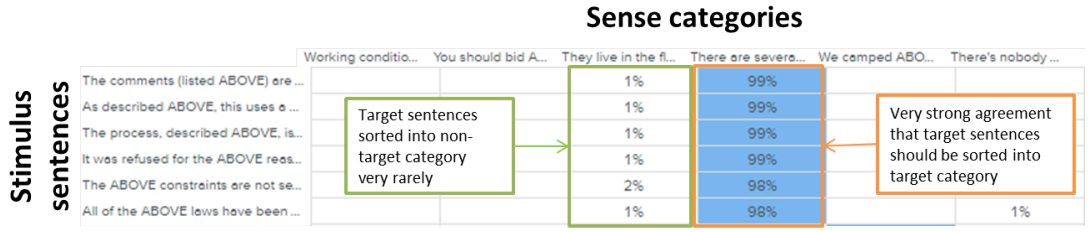


Figure 26 Snapshot of TEXT USE target sentences that have been sorted with a high degree of agreement into the target category.

While the TEXT USE group is very discrete, the picture is somewhat murkier elsewhere. The matrix shows that participants disagree with me and with each other about where certain sentences should be sorted. This is shown in the percentage values displayed outside of the blue-shaded boxes. Recall that these shaded boxes are the groups which have the highest level of agreement as calculated by the OptimalSort grouping algorithm. So while the algorithm has been able to produce discrete groups, some participants sorted members of those groups into other groups.

Participants and I disagree most when deciding which spatial examples share the same meaning of *above*. It is not the case that participants tend to use only one category to encompass all spatial uses, as is evidenced by above-chance values in both of the spatial categories in the popular placement matrix, as shown in Figure 27. This shows disagreement between me and participants, and amongst participants, over how distinctly spatial examples should be categorised.

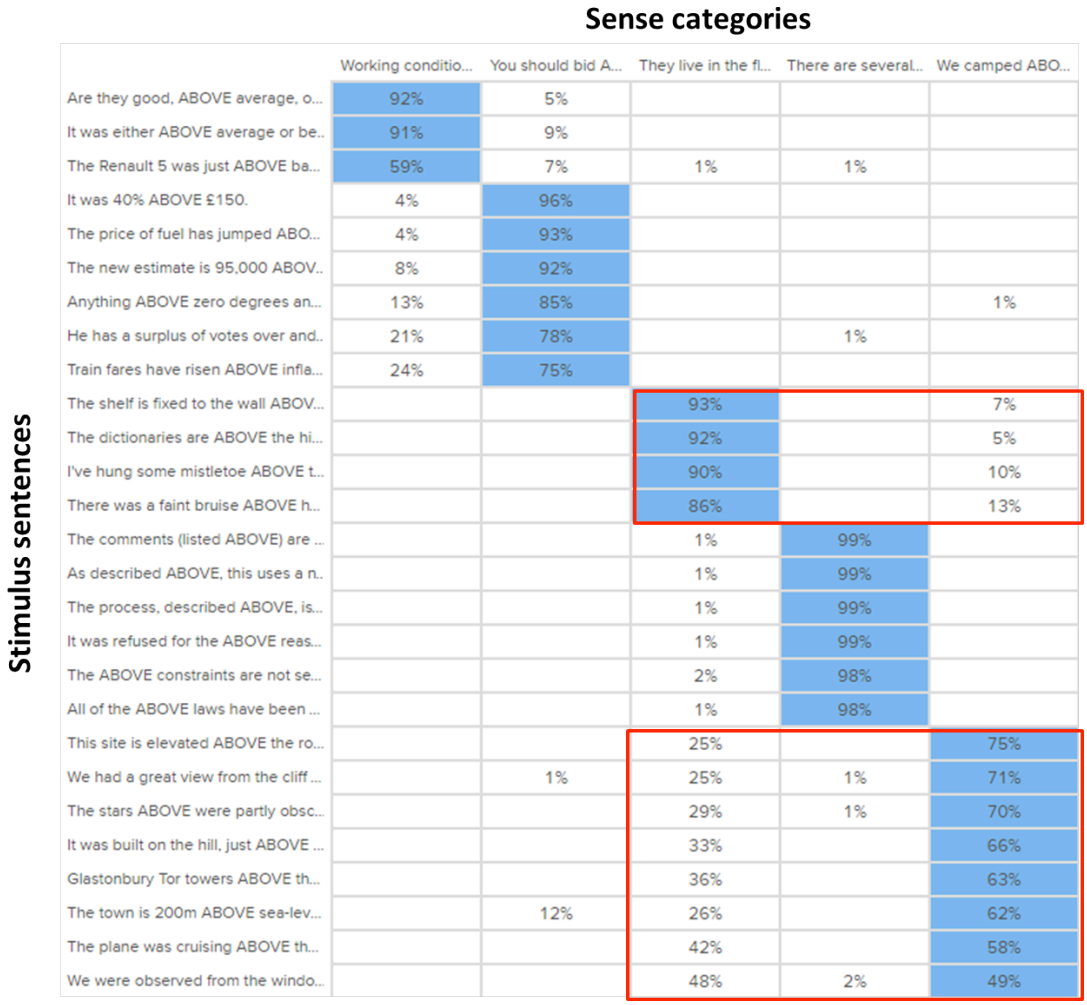


Figure 27 Snapshot of popular placement matrix for *above*, with spatial examples in red boxes

Instead, some participants and I simply differ in our judgments of what sense of *above* each sentence captures. In particular, there is disagreement about how sentences which I judge to be examples of the VANTAGE category should be sorted, with many participants sorting them into the more generic VERTICAL RELATIONSHIP category. While these sorting "errors" are bidirectional, in that target sentences from each are sorted into the other category, there is a greater tendency for target sentences from the VANTAGE category to be sorted into the VERTICAL RELATIONSHIP category than vice versa. This outcome hints at the possibility that while I find the distinction meaningful, not all participants do. Specifically, some seem content that the sentences in the VANTAGE category are not meaningfully distinct from those in the generic VERTICAL RELATIONSHIP category. However, since true lumping, i.e. assigning all spatial examples to one group, is so uncommon (it is done by just six

participants), this disagreement cannot simply be chalked up to my predilection for splitting.

Just as we can observe a bidirectional overlap in the two major spatial categories, there is an overlap in the membership of the non-spatial categories, as shown in the snapshot of the popular placements matrix in Figure 28.

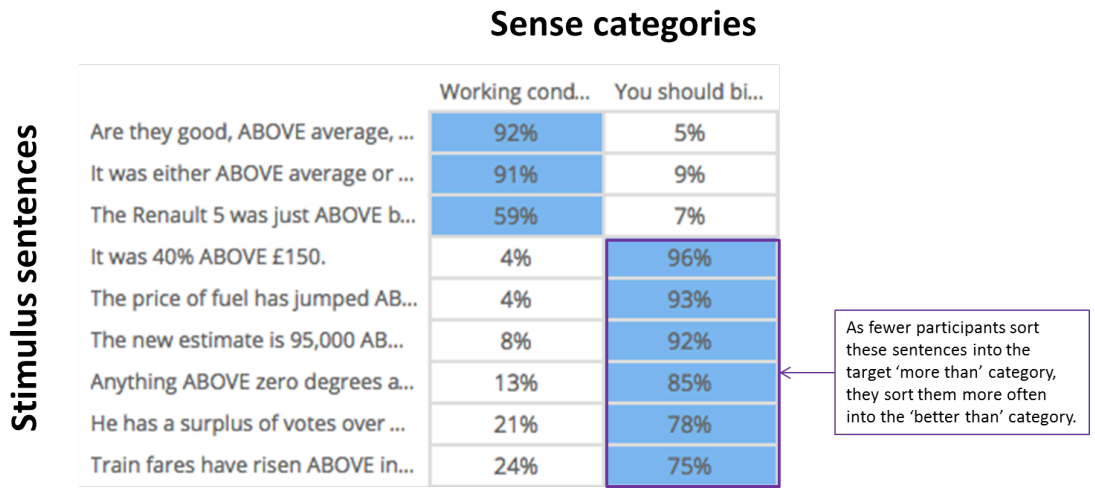


Figure 28 Snapshot highlights tendency for sentences describing quantitative scales

Specifically, this section of the matrix indicates an overlap between examples of *above* to describe qualitative and quantitative scales. Many participants sorted target sentences from the MORE THAN (i.e., quantitatively more than) category into the BETTER THAN (i.e., qualitatively more than) category; the reverse was also true but to a more limited extent. This seems surprising, since if there was to be an overlap between these two groups, I had expected it to operate in the opposite direction. That is, if participants allocated target sentences to the other category, I would have expected examples of *above* that described qualitative scales to be sorted into the category describing quantitative scales, on the basis that it seems more intuitive that qualitative scales are a metaphorical extension of quantitative scales. This finding indicates that some participants and I not only disagree about the meaning of *above* underlying certain senses, but that we also diverge in the way we judge the sentences to be related. Whereas my intuition is that qualitative uses of *above* are an extension of a quantitative sense, the sorting decisions observed here suggest that some participants consider a qualitative sense to be sufficiently generic to also encompass sentences describing quantitative scales, and not vice versa.

There is further overlap between the BETTER THAN and HIERARCHY categories, as shown in the final three rows in the popular placements matrix in Figure 29. While a large minority of participants allocate these final three stimuli to the target BETTER THAN category, most allocate them to the HIERARCHY group. While my classification of these stimuli is different, these participants' decisions are nonetheless cogent and understandable. There is an obvious relationship between superiority in a hierarchy and an individual's sense of personal superiority. For example, the manager of a bank would perhaps be justified in thinking that he is "*above* work like this", if that work was filing papers. I classify all of the BETTER THAN target sentences together, but more than half of the participants separated them and allocated some to a related but distinct category. This clearly suggests a difference between the sense boundaries that I find meaningful and the distinctions made by other native English speakers.

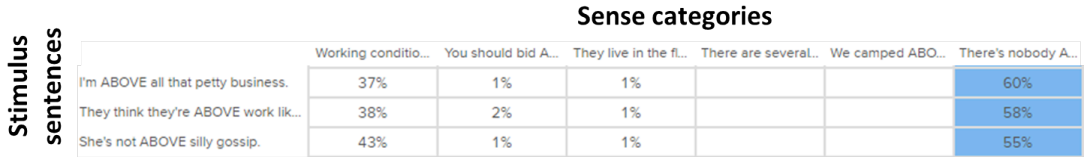


Figure 29 Snapshot of popular placement matrix for *above*, showing overlap between BETTER THAN and HIERARCHY categories

In summary, there are pockets of agreement alongside pockets of murkier, less consistent sorting. That reduction in consistency is not limited to consistency with my sense distinctions, but also reflects poor agreement amongst the participants themselves. The fair kappa values are consistent with this finding.

2.7.2.5 Below

There is widespread overlap in the composition of groups created by participants in this task. This is particularly the case amongst spatial sentences, but there is a lack of clear, discrete grouping amongst non-spatial sentences too. Disagreement over how spatial examples of *below* should be sorted manifests in target sentences being sorted into non-target categories. Figure 30 illustrates the way distinctly spatial examples were sorted, with the green box indicating where these sentences were frequently sorted. The matrix shows that there is considerable disagreement over where distinctly spatial examples of below should be sorted. Further, it shows that while there are some distinct groups, the level of agreement that each sentence belongs in each group is variant.

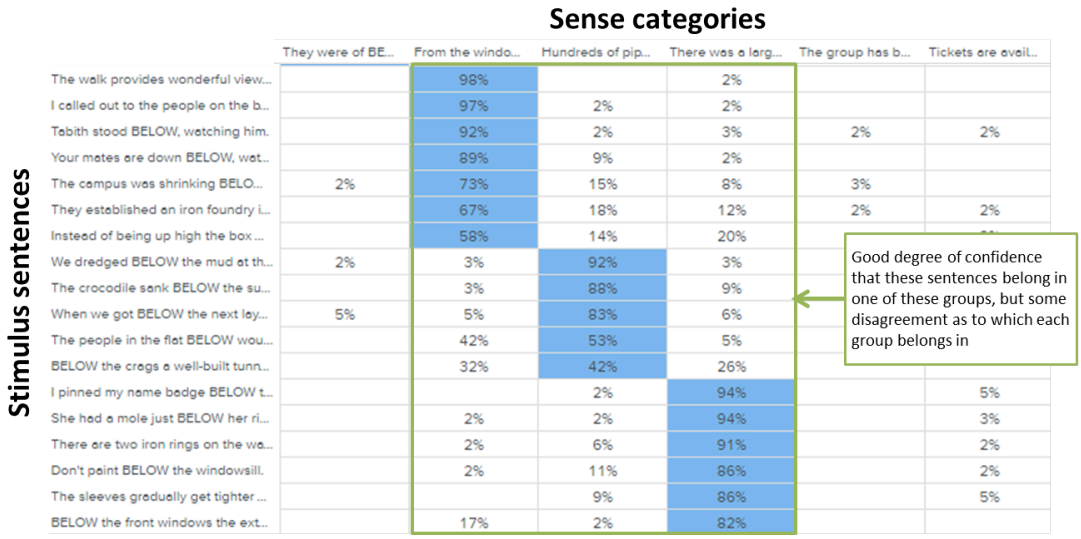


Figure 30 Snapshot of popular placement matrix for *below*

While there are three reasonably distinct groups of spatial sentences evident in the popular placements matrix, there is a considerable degree of overlap across them. This is especially the case with the sentences that are judged by participants to fit best into the UNDERNEATH (3D) category. With the exception of the final example in the group, *Below the crags a well-built tunnel could be seen* (which I judge to be an example of the LOWER THAN (2D) category), the sentences are sorted into the group at least more often than would be expected by chance and, in general, with a good degree of consistency. In the case of this sentence, this data suggests that it is a poor example of any of the three spatial categories given, and that participants may think it should be classified into a different category not provided here. However, across the members of the UNDERNEATH (3D) group, some participants sort the target sentences of this category into another group, and specifically into one of the other two overtly spatial groups. Amongst the other two groups that emerge in the popular placements matrix (VANTAGE and LOWER THAN (2D)) there is further disagreement over where constituent sentences should be sorted, but to a lesser degree. This outcome indicates a considerable divergence between my categorisation decisions and those of the participants and, in addition, amongst the participants as a group. My intuitions about the meaning captured by each sentence, and whether it is equivalent to the meanings underlying the other uses, fail to correspond with all of the participants' intuitions. However, this non-alignment of our intuitions is a matter of degree: there are some instances, for example, *The walk provides wonderful views*

of Mallerstand below, about which participants and I overwhelmingly agree about its meaning.

As we saw in the analysis of the popular placements matrix created for *above*, there is an overlap across the LESS THAN (i.e., quantitatively less than) and WORSE THAN (i.e., qualitatively less than) categories, as shown in Figure 31.

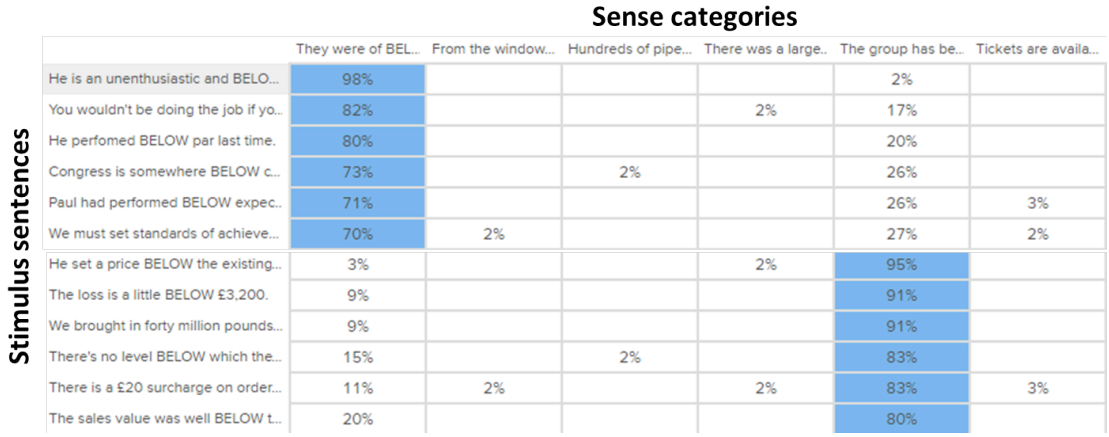


Figure 31 Snapshot of popular placement matrix for *below* showing overlap across LESS THAN and WORSE THAN categories

In this case, though, when participants allocated target sentences from one of these categories into the other, they tended to do this with sentences that were intended to belong to the qualitatively less than category. In other words, if participants did not assign it to the target category, they were more likely to sort sentences describing a position on a qualitative scale into a group capturing positions on quantitative scales than vice versa. This is the opposite of what happened in the *above* sorting task. It seems, then, that the direction of the relationship between qualitative and quantitative scales varies across the two words. This finding indicates that while some participants and I disagree on the meaning captured by these twelve sentences, we do seem to agree on how quantitative and qualitative senses of *below* are related. Although some participants and I disagree about how exactly these stimuli should be sorted, we do agree that quantitative uses are more basic than qualitative uses, as evidenced by the use of the quantitative category to encompass both sense types.

Finally, moving onto the TEXT USE examples of *below*, the matrix reveals that these examples are organised into a fairly discrete group, but there is some overlap with other groups and their constituent sentences. This is illustrated in the final six rows

of the popular placements matrix, which show that target sentences from this category were sorted into other categories by a small number of participants, and in the final column, which shows that non-target sentences were sorted into this group. It is noteworthy that there is a tendency for target sentences, when allocated to non-target categories, to be allocated to groups describing spatial configurations. Similarly, there is a greater tendency for non-target spatial uses of *below* to be sorted into the group than non-spatial uses[3]. This indicates that some participants judge these uses to have a spatial, rather than non-spatial (e.g., temporal) meaning. However, on the whole there was a great degree of consensus about how these particular examples should be sorted, in line with my own intuitions. This discrete grouping indicates that my intuitions about the meaning of *below* in these sentences maps very closely onto those of the participants.

In summary, we have again observed rather a lot of disagreement in how examples of *below* describing both spatial and non-spatial domains are sorted. That said, the groups that do emerge, however tentatively, do seem to align with my intuitions about the senses of *below*. There are exceptions, though. For example, two stimuli, *We dredged below the mud at the bottom of the river* and *Below the front windows the extension was divided into two sections*, were classified more frequently into a non-target group than the target group. This provides a clear indication that some participants and I have different intuitions about what constitute examples of particular senses of *below*. An open sort task will reveal whether this is the result of individual differences in word senses.

2.8 General discussion

In this section I provide a summary of the principal findings of this study. I then address the implications of the findings for our knowledge about word senses, and in particular the practical implications for the role of expert intuitions in the study of polysemy. I then identify some areas of uncertainty that have been raised in this chapter, and close with some concluding remarks.

---

[3] The example *It will be argued below that economic reconstruction was a success* was sorted into the LESS THAN category by 8% of participants. I speculate that this is due not to participants believing that *below* in this example is used to describe a position on a quantitative scale. Instead, I argue that it is because participants have focused more on the presence of "economic success", which invokes notions of finance and consequently quantitative scales, and less on the meaning of *below*.

2.8.1 Summary

I aimed to conduct an empirical investigation of how well naïve speakers' and other linguists' intuitions about word senses correspond to my own. I acted on the assumption that polysemous words (or their senses) are linguistic categories. Accordingly, I assumed that polysemous words (or their senses) would behave like categories; i.e., it would be possible to sort examples of polysemous words into categories according to some categorisation criterion. In this case, the criterion used was the meaning of the polysemous word. I therefore used a categorisation task, operationalized as a sentence-sorting task, to assess whether the way I categorise examples of the polysemous words *over, under, above* and *below* was systematically different from or similar to the way these examples are categorised by other English speakers. If participants and I categorise the sentences in a similar way, we might conclude that in this case, expert intuitions about the senses of polysemous words *do* correspond to those held by other speakers. If it is not the case, it will cast doubt on the representativity of expert intuitions about word senses.

A set of four closed sentence-sorting experiments, each completed by native English speakers, was used to answer this question. The magnitude of agreement between me and each participant was calculated statistically using Cohen's kappa. This quantitative analysis reveals that agreement between me and each participant varies across participants. For example, the data suggest that within participants who sorted examples of *over*, some participants and I reached identical conclusions about how the stimuli should be categorised. In contrast, some participants and I reached much weaker consensus. Further variation was observed across the four words. Of the four words studied, participants in the *over* task reached the strongest degree of agreement with me, whereas participants who sorted examples of *under* reached weakest agreement with me. Further still, variation in agreement with how particular sentences should be categorised was observed. For example, as noted in section 2.7.2.5, analysis of the popular placement matrices showed that there were pockets of good agreement, resulting in clearly delineated categories. However, these were interspersed with pockets of poor agreement, reflecting disagreement not only with how I classify the sentences, but with how other participants classify them. Differences in the way participants and I categorise examples of these words indicate that we differ in what sense we judge each sentence to exemplify. A qualitative

analysis complemented this quantitative analysis, and identified particular example sentences that participants and I sorted in markedly similar or different ways.

One of the most interesting outcomes of this study is that agreement with my sense distinctions was not a binary affair; instead, agreement was really quite variant. This is interesting because it suggests that some participants, with whom I had high agreement levels, may agree with me about what the senses of these words are, whereas others, with whom I had low agreement level, may have different senses to mine. Within those two extremes, agreement values varied. The fact that some participants and I may share senses, whereas some may have different senses to me, is an early indication of individual differences in word senses.

2.8.2 The role of linguists' intuitions in the study of polysemy

Despite the highly structured nature of these tasks, we observe a range of agreement values across individuals. These findings suggest that the representativity of the senses I find meaningful in the selection of examples of each target word in the minds of other speakers is a matter of degree; it is neither the case that participants and I completely agree about what senses the example sentences represent, nor is it the case that my intuitions about the boundaries of the senses of these words do not map onto those held by other speakers at all. This finding – quantified through statistical analysis – adds an extra detail to the literature on the compatibility of expert and naïve speakers' intuitions about linguistic phenomena. Specifically, it complements research in syntax which has also found conflict between expert and naïve intuitions. The qualitative analyses provide additional information about where my intuitions converge and diverge with those of a large sample of native speakers of English.

It is interesting that of the four words, *over* is the one in which participants and I reached highest levels of agreement. This is despite the fact that *over* is one of the most closely-studied words in the cognitive linguistic literature. While my study of the four words is likely to be fairly evenly distributed, it is reasonable to expect that my exposure to theoretical studies of *over* and the sense distinctions posited in publications may have skewed and distorted my "true" sense distinctions. *Under, above* and *below* have received far less attention (Evans and Tyler, 2005 present a

limited exploration of these prepositions and *over*, but I am aware of no other published study of the polysemy of these three words). I would anticipate that this may make the sense distinctions at which I arrive truer, and less influenced by those proposed in the cognitive linguistic canon. In the context of the debate over the impact of close study of particular linguistic phenomena on a linguists' intuitions (cf. Snyder 2000; Spencer 1973), I would predict that the distinctions I find meaningful in the examples of *under, above* and *below* would be most consistent with those of the participants. Instead, the opposite appears to be true. This outcome complicates the claim I have just mentioned, and serves to support the interpretation that focused attention on a polysemous word, instead of resulting in an unrealistic conception of its senses, actually coincides with intuitions which correspond to those of other native speakers. *Over* is quite unusual in the extent to which it has been studied in cognitive linguistics. In the absence of such intense focus on the polysemy of other words, it is difficult to test whether this is an outcome unique to *over*. It is therefore not possible to reach a firm conclusion about this, but it remains an interesting possibility, and one which encourages further close studies of other polysemous words.

2.8.3 Lumping and splitting

One of the key objections made in literature critical of cognitive linguistic accounts of polysemy is the finely-grained sense distinctions certain linguists propose (e.g., Tyler and Evans 2001, p. 761; Sandra and Rice 1995). This study indirectly tests whether this criticism is warranted. The data show that, on the one hand, a small minority of participants sorted the stimulus sentences into groups in a way that was very similar to my classification decisions. On the other, there was a large proportion of participants who sorted the cards rather differently to me. Inspection of the decisions participants at the high and low ends of the agreement scales made about whether examples should be "lumped" or "split" reveals predictable results at the high end: all participants with agreement scores of 0.9 or higher used all six preset groups. At the other end of the scale, looking at participants with whom I reached an agreement level of up to 0.6, the outcome was less predictable. Rather than being able to explain the weak agreement scores on these participants' decisions to use a smaller number of groups, and therefore adopt a lumping approach that literature on the danger of linguists' sense distinctions implies that naïve participants might take,

most participants at this end of the scale used all six groups, too. Of the nineteen participants with whom I had an agreement score below 0.6, only seven used fewer than six of the preset groups.

On the whole, participants used an average of just under six groups to sort the sentences in each task. This suggests that the number of distinctions I presented in the task, and the number that I find to be meaningful within the stimuli, corresponds well to the number of distinctions that participants find meaningful. In the face of this consistency, we must seek an alternative explanation of the scarcity of very high agreement values. A possible explanation for this is that while participants recognised that there were six meaningful distinctions within the stimuli, some of the stimuli they believed to exemplify those distinctions were different to those which I believe to be exemplary. The close analysis given in section 2.7.2 confirms this.

2.8.4 Linguists' agreement with my intuitions

Table 8, which shows kappa values of agreement between me and participants classified as linguists, contrasted with Table 7, containing agreement values between me and all non-linguist participants, reveals some interesting outcomes. It is not the case that linguists and I consistently reach higher levels of agreement. In fact, the opposite is true. In three of the four target words, non-linguists and I typically agree more than linguists and I do; *below* being the exception. Further, non-linguist participants have higher maximum kappa values in three of the four words, and in the case of *under*, the highest kappa value is the same as the highest kappa value scored by a linguist participant. However, the minimum kappa value between me and non-linguists is always lower than between me and linguists. On the whole, the data point to an unexpected conclusion that non-linguists and I tend to agree more about how the sentences should be categorised than linguists and I do. It should be noted, however, that the small number of linguists sampled makes this a tentative conclusion, but it is an interesting finding that warrants further study.

These findings add further detail to the debate over the utility of linguists' intuitions in the analysis of a particular linguistic phenomenon. In this case, the data offer an early suggestion that a group of participants with advanced knowledge of linguistics do not share a set of senses of *over, under, above* and *below*. This adds weight to the

argument against reliance on introspection as a methodology. If one linguist's intuitions do not correspond to those of another linguist's – and, as shown here, with multiple other linguists – what confidence can we have in the notion that intuition-led analyses capture the reality of the particular phenomenon in the language and/or minds of other people?

2.9 Questions raised

It is clear that participants and I did not consistently and reliably agree with each other about whether a particular example of *over, under, above* and *below* constitutes an example of a particular sense. This suggests that my intuitions about the senses of these words may not neatly correspond to those of other speakers. However, this study cannot reveal whether the participants in my study share a set of senses of these words, and it is simply the case that those senses are different to mine, or whether individuals have different senses of these words. However, the fact that agreement values were so variant suggests that there may be individual differences in word senses. A variation on the task used in this study, in which participants are not given predetermined categories but are free to create their own, can shed light on this possibility. The next chapter describes a study that explicitly aims to understand whether the outcome found here is due to a mismatch between my senses and a set common to other speakers, or due to individual differences in word senses.

2.10 Conclusions

Analysis of the sorting decisions made by participants using both quantitative and qualitative approaches has revealed that there are tendencies to both agree with the sense distinctions I find meaningful within task stimuli, and disagree with them. Across the four words studied, there were cases in which participants sorted stimuli in very similar ways to me and to each other; this was particularly the case in the *over* task. But equally, there were cases in which participants disagreed with me, and with each other, over where some stimuli should be sorted. In the most extreme cases, participants agreed with each other more than they did with me, with the majority choosing to assign a sentence to a non-target category. The data presented here, analysed quantitatively and qualitatively, indicate that my intuitions about what the senses of four polysemous words are do not consistently and reliably correspond to those held by other native speakers of English, regardless of whether or not they too are linguistics experts. The aim of this chapter, to study the correspondence

between my intuitions and those held by others, has therefore been met, and the finding made here is an original contribution to knowledge.

Some linguists (e.g., Tyler and Evans 2001, p. 761; Sandra and Rice 1995) have levelled criticism at highly fine-grained approaches to analysing the senses of a given polysemous word, therefore implicitly questioning whether finely-grained analyses are psychologically realistic. I judge that my intuitions about how the stimuli used in these tasks should be divided betray a tendency towards a fine-grained analysis. In the face of criticism about finely-grained analyses, it is therefore interesting to note that participants used an average of just fewer than six groups to categorise the stimuli. Given that I divided the stimuli into six groups, it appears that participants and I, on the whole, have remarkably similar intuitions about what level of sense distinction within the stimuli is meaningful. This finding, in contrast with the variation in how well individual participants and I agreed about how the stimuli should be categorised, suggests that future critiques of intuition-led analyses of polysemous words should not be overly concerned with whether or not finely-grained approaches are cognitively realistic – they appear to be – but about how well another speaker would agree that a given sentence is an example of a particular sense.

At this point it can be concluded that participants and I differ in what constitutes examples of six senses of *over, under, above* and *below.* It is not possible to determine whether the senses held by participants are homogenous, and it is simply the case that I differ with them, or whether individuals have different senses of these words. However, given that participants (dis)agreed with me about how the sentences should be sorted to different degrees, it seems possible that this is due to the fact that some participants and I agree more about what the senses of these words are than others. This indicates individual differences in word senses, but needs further and more direct study.

Uncovering variation in language is the source of interesting facts (Schütze 1996, p. 9). The results of this sorting experiment support this idea, and has not only indicated that my senses of *over, under, above* and *below* do not always coincide with those of other speakers, but also hinted that there may be individual differences

in word senses more generally. The following chapter, which describes open sentence-sorting tasks, directly addresses this possibility.

Chapter 3 Experiment 2: An open sentence-sorting task to test for individual differences in word senses

Data collected in the closed-sort tasks reported in Chapter 2 indicate that participants did not consistently agree with the sense distinctions I found to be meaningful in the uses of *over*, *under*, *above* and *below* used in those tasks. Moreover, disagreement with my sense distinctions, as manifested in the differences in the way participants and I sorted the stimuli, varied across participants, and across the four words tested. This first finding can be interpreted in two ways. First, it could be that my participants share a set of senses of the four words tested, and these senses are different to mine. Under this interpretation, variation might be explained on the basis that some participants were able to exploit the structured nature of the task to infer how the sentences "should" be sorted, rather than sorting them according to their own intuitions about similarities and differences in the meanings of the target words. The second explanation is that participants have different senses of these four words. The notion that there may be individual differences in word senses should not be particularly controversial, since it has been demonstrated that there are individual differences in other areas of language. However, studies of polysemy, and indeed dictionaries, implicitly assume no variation in word senses across individuals. This chapter explores the possibility that there are individual differences in word senses, and consists of two parts. The first part consists of a review of relevant literature that has addressed individual differences in language in general, and which has started to address, albeit indirectly, individual differences in word senses in particular. The second part reports an open sentence-sorting task that aims to study whether there are individual differences in word senses. As noted in Chapter 2, this thesis assumes that polysemous words are linguistic categories, and that they should therefore behave like categories. In other words, it should be possible to organise examples of polysemous words into groups according to some categorisation criterion. In this case, the criterion of interest is the meaning of the polysemous word. The chapter

closes by discussing the implications of the findings made, and identifying areas for further investigation.

Part 1: Literature review

3.1 Individual differences

The proposition that not all members of a linguistic community acquire the same grammar is one that runs counter to mainstream, generative assumptions (both implicit and explicit) about linguistic competence. For example, the principles and parameters account assumes that humans are genetically endowed with principles (i.e., highly abstract universal rules), and that parameters are "switches" specifying the range of permissible variation within the principles. It is argued, for example, that X' phrase structure, which states that an X-phrase (e.g., NP) will always contain an X (e.g., N) as its head. This is a principle that is argued to be genetically endowed. The *position* of the head in the phrase, however, is a parameter that varies across languages. For example, English is a head-initial language, whereas Japanese is a head-final language. It is claimed that exposure to the target language sets these parameters. Accordingly, it is assumed that all speakers in a given language community will ultimately acquire the same grammar.

It has been demonstrated, however, that speakers within the same language community do not appear to acquire the same grammar, and that there is evidence of significant individual differences in language acquisition. Differences across individual children are not limited to particular areas of linguistic competence; they have been recorded in early phonology (Leonard 1980), joint-attention skills (Mundy and Gomes, 1998), lexical processing (Fernald and Marchman, 2012), the paths followed to reach multi-word speech (Pine and Lieven, 1993; Shore, 1995), vocabulary (Bates et al. 1995), syntax (Vasilyeva, Waterfall, and Huttenlocher, 2008), and semantics (Rice 2003).

Research in this area has also considered individual differences in language attainment amongst adults. In her review of a number of investigations of differences in linguistic attainment between high academically achieving and low academically achieving populations, Dąbrowska (2012) notes that while the more educated

participants typically performed at ceiling level – meaning that their responses are typically homogenous – participants who had completed less education showed much more variation in their responses to a range of linguistic stimuli. More recently, Divjak, Dąbrowska, and Arppe (2016) have argued that there is evidence that a single model of grammar cannot account for all attested usage of English, but that instead a very large number of grammars are compatible with actual use.

Beyond individual differences in language acquisition and grammatical attainment, it has also been demonstrated that there are individual differences in the interpretation of ambiguous temporal metaphors. Duffy, Feist, and McCarthy (2014), for example, have observed differences in interpretation of McGlone and Harding's (1998) ambiguous statement *The meeting originally scheduled for next Wednesday has been moved forward two days*, noting that individuals with different personality types interpret the statement differently.

3.2 Individual differences in polysemy and word senses

While individual differences in language acquisition, adult attainment and beyond have firmly been firmly established in cognitive linguistic literature, scholars in the field have not, to my knowledge, studied the possibility that there may be individual differences in polysemy and word senses. This is despite the fact that one model of categorisation that has been used to account for polysemy, the exemplar model, implicitly predicts individual differences in categorisation decisions and categories. It has, however, been observed, albeit indirectly, in computational linguistics research. Specifically, scholars in word sense disambiguation research have found that agreement with "gold standards", i.e., with expert judgments about which sense tag should be applied to each individual example of a polysemous word, varies across individuals.

When asked to assign examples of a particular word into a predetermined category based on the meaning of that word, individual annotators agree with the gold standard to varying extents. An example of exceptionally high inter-annotator agreement is given in Snow et al.'s (2008) study, in which individuals recruited through Amazon Mechanical Turk were asked to classify examples of the word *president*. Annotators reached 100% consensus with the gold standard, though it

must be stated that this noun was to be classified into one of only three possible sense categories. An example at the opposite end of the scale comes from Passonneau, Baker, Fellbaum, and Ide's (2012) WSD study, which, when asked to use preset tags to annotate examples of *normal,* annotators agreed with each other worse than would be expected by chance. Within these two extremes. studies have demonstrated that more intermediate agreement is visible amongst annotators. For example, Bhardwaj, Passonneau, Salleb-Aouissi, and Ide (2010) found that agreement amongst six annotators ranged from "about 0.5 to 0.7 for nouns and adjectives, and about 0.37 to 0.46 for verbs" (p. 4), and Passonneau, Salleb-Aoussi, Bhardwaj, and Ide (2010) found that six annotators' decisions gave agreement scores ranging from 0.37 to 0.68. The fact that in these and similar tasks the annotators each undertake identical tasks indicates that there are differences in the extent to which individuals agree with each other in how the preset sense categories are used, and therefore they vary in how well they agree with the gold standard, i.e., with expert sense distinctions. Further, the extent of their agreement varies according to the word they are examining (Bhardwaj et al., 2010; Passonneau, Bhardwaj, Salleb-Aouissi, and Ide, 2012; Passonneau, Salleb-Aouissi, and Ide, 2009). In an effort to understand why the same individuals agree with the sense distinctions of some words better than they do others, it is suggested that a number of factors may be at play. Passonneau et al. (2009, p. 3) set out three factors, in addition to individual differences amongst annotators, which they judge to have some impact on agreement values, all of which seem reasonable and intuitive:

> "1. Greater specificity in the contexts of use leads to higher agreement
> 2. More concrete senses give rise to higher agreement
> 3. A sense inventory with closely-related senses (e.g., relatively lower average inter-sense similarity scores) gives rise to lower agreement."

In later work, the authors further explore the effect individual words have on inter-annotator agreement, and suggest that words that can be found in contexts that are more open to subjective interpretation may explain some cases of disagreement; this likely ties in with the second factor, cited above. They suggest, for example, that the adjective *fair* is inherently more subjective than the adjective *long*, which describes a measurable physical quality. A later paper sharing some of the same authors also notes that the subjectivity of a word may be a product of different "perception[s] and

experience[s] of individuals", and give *justice* as an example of a word that derives its "meaning from cultural norms that may differ from community to community" (Bhardwaj et al., 2010, p. 2).

These studies provide clear evidence in support of the notion that different people assign the same examples of a word to different sense categories, which is an early indication of individual differences in word senses. To a reader well versed in natural language processing literature, these findings will not come as a surprise. As Passonneau et al. (2009, p. 3) put it, "It is widely recognized that achieving high K [agreement] scores (or percent agreement between annotators) [...] is difficult for word sense annotation." It is later acknowledged that "variation in word sense annotation across annotators should be expected as a consequence of usage variation" (Passonneau et al., 2010, p. 1). Later still, it is noted that "In the lexicographic and linguistic literature, it is taken for granted that there will be differences in judgment across language users regarding word sense" (Passonneau, Bhardwaj, et al., 2012, p. 11). If these differences are taken for granted in the cognitive linguistic literature, this is not an assumption that is widely acknowledged; indeed, typical studies of polysemous words make no reference to the possibility that different speakers would judge a single example of a target word to be an example of different senses. It is clear, then, that in some branches of linguistic study word senses are understood as not being "discrete, atomic units that can be delimited and enumerated" (Bhardwaj et al., 2010, p. 2; cf. Erk, McCarthy, and Gaylord, 2009). Moreover, Bhardwaj et al. (2010, p. 2) believe that the division of word senses is a "somewhat artificial" enterprise, which in itself may explain why perfect agreement in human WSD is so unusual. Even the standards by which those orchestrating WSD studies should delineate the senses to be offered to annotators is the subject of disagreement (Passonneau, Bhardwaj, et al. 2012); typically annotators can use only one sense label, despite evidence indicating that sense annotation is not necessarily an all-or-nothing affair, and that sense tags are graded in their applicability to example sentences (Erk, McCarthy, and Gaylord, 2009, p. 17).

While these studies provide further evidence, consistent with the findings of Experiment 1, that there may indeed be individual differences in word senses, they primarily contribute to the ongoing debate over the traditional role of introspection

and authors' intuitions in cognitive linguistic research, and its role in the study of polysemy in particular. This link is perhaps best articulated by Carletta (1996, p. 1) in her research into measuring agreement using the Kappa statistic, in which she states that "researchers are beginning to require evidence that people besides the authors themselves can understand and make the judgments underlying the research reliably[, and that] if researchers can't even show that different people can agree about the judgments on which their research is based, then there is no chance of replicating the research results." The research presented here provides a rather stark insight into the problematic nature of positing word senses.

3.2.1 Effective communication in the face of individual differences in word senses

If individuals do have different senses of polysemous words, how is it that we manage to communicate effectively? A possible explanation has been offered by computational linguists Ide and Wilks (2007). They propose that humans disambiguate polysemous words by accessing a general, coarse sense which is progressively distinguished to isolate a sub-sense only if that sub-sense is needed for understanding. They further propose that human language comprehension can generally operate at the level of homograph. While this suggestion was not offered in the context of individual differences in word senses, it does offer a possible explanation of how humans can communicate in the face of such differences, should they exist. It seems feasible that, if individuals differ in how they categorise a set of sentences, as long as they agree that they can be collapsed into a broad, albeit perhaps less meaningful sense, they can understand each other. In that way, where individuals' word senses don't completely overlap, a "good enough" level of understanding can be achieved by accessing this coarse sense.

3.3 Individual differences in word senses: conclusions

Despite the fact that interest in and awareness of individual differences in language is growing, and that polysemy remains a topic of ongoing debate, research in cognitive linguistics has shown little engagement with the possibility that different people may have different senses of a polysemous word. Evidence from WSD research using a small number of human annotators suggests that there may be some inter-speaker variation in word sense boundaries. While larger-scale WSD research – research that depends on the crowd to undertake human WSD – is ongoing, to date it

has not adopted open-sort methods, which would see annotators creating their own sense categories for each word. This would give a fuller picture of the extent to which individuals' senses overlap, and is the methodology the study presented in this chapter uses to address this particular issue. Until that point, poor inter-annotator agreement values in WSD studies can only *indirectly* indicate individual differences in word senses.

If individual differences *do* exist in word senses, this would provide some support for a theoretical model of their representation. The radial models proposed by Brugman and Lakoff (2006 [1988]) and Tyler and Evans (2001) do not leave open the possibility that individuals may have different senses of a given word. In fact, the fact that they claim that *over*, for example, has a particular set of senses in itself suggests that the authors posit a model of word senses in which there are no individual differences. In contrast, a model based on exemplar theory can accommodate individual differences in word senses, on the grounds that categories are understood to comprise of tokens of previously-encountered exemplars that are organised according to the demands of a particular categorisation event. Where exposure to a given polysemous word varies, or where the exact characteristics that one attends to in a categorisation event vary, it seems conceivable that the linguistic categories speakers construct may also vary.

The literature presented here motivate a study that improves upon the methodology adopted by scholars to date. The limitations of using closed-sort tasks, analogous to sense-tagging exercises used in WSD research, to study individual differences in word senses were identified in Chapter 2; in brief, a closed sort task does not reveal whether differences in levels of agreement with expert sense distinctions is the result of individual differences, or the result of all participants having a homogenous set of senses which differ to those of the expert, and it is simply the case that some participants used the sense boundaries indicated by the preset categories to guess at a "correct" sorting solution. The findings reported in Chapter 2, along with research to date on individual differences in language and word senses, motivate a direct study of differences in word senses. Specifically, it encourages the use of an open-sort task, in which participants create their own categories.

Part 2: Investigation

3.4 Aims

The aim of the study described in this chapter is to further investigate the possible explanation for the data presented in Chapter 2. That data allowed the interpretation that participants' disagreement with my sense distinctions may be the result of individual differences in word senses. Since it requires participants to use a predetermined set of categories, a closed-sort task like the one used in that study cannot fully reveal the distinctions that individuals find meaningful. Only a task in which participants can construct their own categories, such as an open-sort task, can do this. This chapter therefore reports an open-sort task designed to specifically address the possibility that there are individual differences in word senses. If there are such differences, I anticipate that these would manifest themselves in the data in the form of imperfect inter-participant agreement scores, and qualitatively different categorisation behaviours.

3.5 Data collection

In this section, I outline the approach I took to gathering data to answer the question of whether there is evidence of individual differences in word senses. It then provides relevant information about the participants who completed the tasks. Thirdly, details of the stimuli used are provided, followed by detailed information about the task procedure. Finally, it introduces the statistical model and visualisation tool used to analyse the degree to which pairs of participants agreed about how the sentences should be sorted.

3.5.1 Methodology

A variation on the sentence-sorting task described in Chapter 2 was used in this study. Specifically, this study uses an open-sort task. In an open-sort task participants are simply given a set of stimulus sentences and are asked to sort them into groups of their own making. In this case, participants are expected to create groups on the basis of commonality of meaning of a capitalised target word. This task therefore differs from closed-sort tasks in that participants are free to choose their own groups, and are not distracted by the presence of preset groups. The groups participants create are understood to reflect individuals' senses.

The value of open sort tasks – that they reveal individuals' senses – comes at some cost. The lack of preset groups means that participants are required to think very closely about the meaning of a particular word in each sentence and decide whether a small variation in the meaning of a target word across two sentences justifies creating a new group. So while at face value the task seems quite straightforward, what participants are asked to do is actually quite demanding, and requires extended concentration while they engage in an unfamiliar activity. For this reason, it is important that the number of sentences given to participants is decided by balancing the need to provide a fair sample of the range of uses of the target word and the need to prevent the task from causing fatigue, which might be realised through sorting decisions that become decreasingly consistent as the task progresses. A related issue that encourages limiting the number of sentences each participant is given is semantic satiation (James 1962). Miller (1971) suggests that participants can sort up to 100 sentences at a time. Research using this methodology has used a range of quantities of stimulus sentences, from nine (Divjak and Gries, 2008) to 244 (Baker 1999); other studies have used quantities that are perhaps more manageable but which offer some degree of variation (e.g., Rice et al., 1999; Rice, 1996, both of which saw participants sorting 50 examples of a given word). The present study sees participants sorting 100 sentences, each containing a particular target word.

Most participants completed the task online, using the OptimalSort web-based sorting tool described in section 2.6.4. Two participants completed the task using printed cards, which were shuffled before the task commenced.

3.5.2 Participants

A total of 44 adult participants completed the task. Eight completed the task for *over* and *above*, seven for *under*, and 21 for *below*. All participants were aged 18 or older and spoke English as their first language. All had completed at least secondary education. The participants did not receive a reward for completing the task. Participants were recruited via word of mouth and Reddit Sample Size.

The *below* task was undertaken after the other sorting tasks. A larger sample was recruited to allow me to assess whether the generally weak inter-participant

agreement observed in the first three tasks could be explained on the basis of the smaller sample size in these groups.

3.5.3 Stimuli

The task stimuli consisted of concordance lines extracted from the spoken and written sections of the British National Corpus (BNC) and transcripts of English-speaking children's language on CHILDES (MacWhinney 2000). Additional sentences, used to add diversity to the sample extracted from corpora, were generated by the author and naïve contributors. Participants were given 100 examples of each target word.

Concordance lines were selected as stimuli on a partially-random basis. A random sample was extracted from the BNC. Due to the exploratory nature of this study, the stimulus concordance lines were selected to represent as broad a range of uses as possible. This entailed that where the sample included an excessive number of instances of what I judged to be the same sense, some of these examples were excluded. The stimuli were balanced for spatial and non-spatial uses.

3.5.4 Procedure

Participants received written instructions explaining how they should complete the sorting task. These instructions are presented in Appendix 3. They were asked to read through them and ask any questions if they were uncertain about the task. They could refer back to the instructions at any point.

Participants were instructed to read each example sentence and focus on the capitalised target word to try to understand exactly what it meant in each context. They were told that they should put examples of each word into groups according to the meaning of the target word. They were told that the goal of the task was to create groups in which each member should have exactly the same meaning of the target word as every other member of the group. Once they had completed grouping the examples, participants were asked to label each group to describe, define or paraphrase the meaning of the target word. Until they confirmed that they had finished the task, participants were free to change the composition of the groups, moving examples in and out until they were satisfied with their groups.

Participants who completed the task in person sorted cards sized 5" x 3", on which each example was printed in black ink. Participants who completed the task online used the web-based OptimalSort software described in section 2.6.4.1. Unlike in the experiment reported in Chapter 2, however, participants in this task were not presented with preset categories. In this case, therefore, sentences were presented in a column on the left side of the screen, with a large, blank sorting pane to the centre and right of the screen. Participants were advised to read all of the sentences, considering carefully what the target word, which was shown in full capitals, meant in each sentence. Afterwards, they were required to move each sentence into the sorting pane to create a group, after which further sentences could be dragged and dropped into it. Sentences could be dragged in and out of categories until the participant was satisfied with their sorting decisions. Once they were satisfied with their groups and had given each a label, they clicked a button to indicate that they had finished the tasks. The results were then stored and accessible in the back end of the programme. Participants were required to sort all of the stimuli before they could submit their responses.

3.5.5 Statistical analysis

Data collected in this task were statistically analysed for agreement using Morey and Agresti's adjusted Rand (Morey and Agresti, 1984). Unlike Cohen's kappa, the agreement statistic used in the closed-sort task, Morey and Agresti's adjusted Rand is used to understand the level of agreement amongst pairs of participants who complete an open-sort task, i.e., when they are not given a set of groups into which stimuli are to be sorted. It does not require subjects to use the same number of groups. Like Cohen's kappa, Morey and Agresti's adjusted Rand is a magnitude statistic, and does not return significance values. Possible agreement values range from -1.0, representing total disagreement, through 0, reflecting chance agreement, to 1.0, representing perfect agreement.

3.5.5.1 Interpreting the data

Before we begin to look at the data generated in these tasks, I will explain how the data were analysed. Morey and Agresti's adjusted Rand was used to calculate the degree of agreement between pairs of participants. A more qualitative approach was taken to analyse the similarity matrices produced, though they do offer some quantitative data. Similarity matrices record the percentage frequency with which

each pair of stimuli is sorted into the same group. Frequencies in the similarity matrices are therefore related to agreement: when all participants sort two stimuli into the same group, represented by a score of 100, they have reached complete agreement about how those two particular stimuli should be sorted. The OptimalSort algorithm produces similarity matrices automatically. Helpfully, it colour-codes the chart and uses increasingly dark tones to represent increasing frequency of pairing. Moreover, it organises the matrix such that pairs of sentences which are sorted together with a high degree of frequency are positioned next to or very close to each other. These two useful characteristics combine to produce a similarity matrix which can reveal groups of stimuli that participants typically sorted into the same group, and shows which pairs of sentences are rarely or never sorted into the same group. The matrices therefore reveal perceived similarity and difference in the meaning of the target word in each pair of sentences.

Figure 32 below shows an annotated simplified example of a similarity matrix. The darker shades of green reveal three groups. In the first group are stimuli 1 to 3, in the second stimuli 4 to 7, and in the final group, stimuli 8 to 10. The first group has the highest degree of agreement, with each stimulus being put into the same group as the other two stimuli by all participants. The second group has a slightly weaker degree of agreement, with one member being paired with another member by 75% of the participants. The final group has a weaker level of agreement still. However, because stimuli 8, 9 and 10 are only rarely sorted into the same group as stimuli 1–7, they can be considered a distinct group.

The lighter shades in the table suggest that there is some degree of overlap in membership across the three groups. Overlap is meant here to describe the fact that members of the strongest groups identified in the similarity matrix are judged by some participants to be members of other groups. This is shown in the percentage values outside of darkly-shaded triangles. For example, in Figure 29, stimulus 1 was sorted into the same group as stimulus 5 by 15% of the participants, and into the same group as stimulus 9 by 25% of the participants. However, because stimuli 1, 2 and 3 were sorted together 100% of the time, and only much more infrequently with any of the other stimuli, they are considered to be a distinct group.

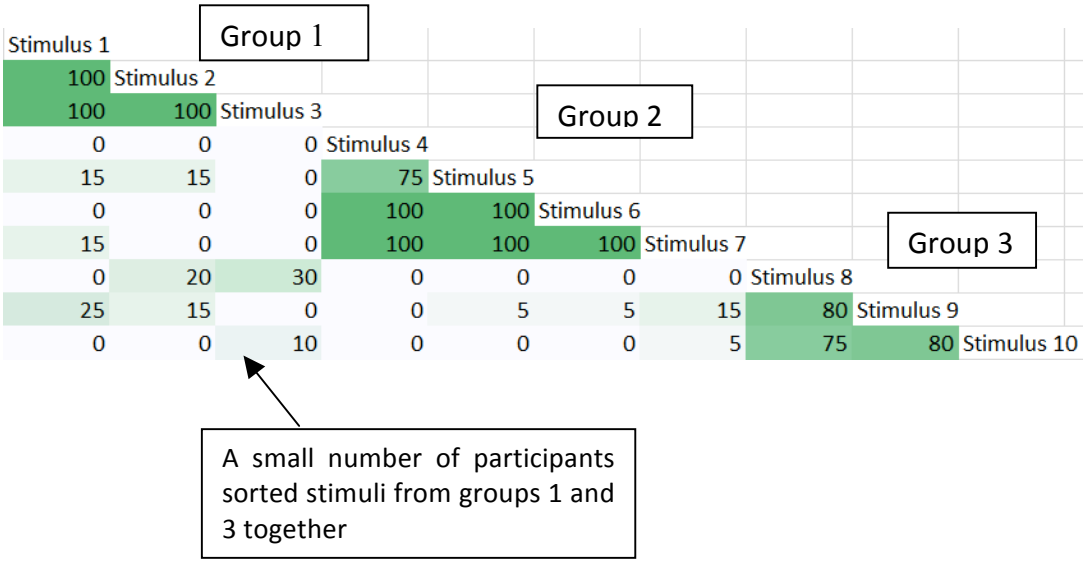


Figure 32 Simplified annotated example of a similarity matrix showing three groups

The groups discussed in the following sections are those which have a good degree of agreement.

So what can similarity matrices tell us about individual differences in word senses? If a similar matrix consists of groups of stimuli – in the context of this study, sentences – which have 100% agreement, that would indicate that the senses represented by the stimuli are agreed upon by all participants. If, as is in the case of the similarity matrix in Figure 29, there are cases where some pairs sentences are judged to use the target word in the same sense by some but not all participants, that would suggest that word senses are subject to individual differences. In this way, similarity matrices therefore allow an additional means of studying inter-participant agreement, complementary to agreement statistics.

3.6 Results and discussion

This section presents statistical and qualitative analyses performed on the sentence sorting data. It begins by presenting agreement values, showing how well participants agreed with each other. I then offer a qualitative analysis of the resultant similarity matrices. This qualitative analysis describes groups of sentences which are often grouped together, and includes my speculation on what the overarching meaning of those groups might be. Not all groups are as readily interpretable as each other, and for that reason I emphasise which groups are clear, and which are trickier to understand. The purpose of this is to establish whether there are particular groups

of sentences that are subject to good agreement. If there are, these may be tentatively understood as word senses that multiple speakers find meaningful. If, on the other hand, there is disagreement about how particular sentences should be sorted, i.e., some participants sort sentence *x* into the same category as sentences *w, y* and *z*, others categorise *x* into a category with sentences *y, a,* and *b*, while other sill categorise *x* with *b, c* and *d*, that would suggest that there is disagreement over what participants judge the target word in *x* to mean. We can tentatively interpret this outcome as evidence of individual differences in word meaning.

I also study whether participants had a tendency to "lump" or "split" the stimuli into coarse or fine groups. This will shed some light on whether or not participants qualitatively differ in the level of granularity they find meaningful. In addition, I will identify whether or not participants create any "mixed" groups containing both spatial and non-spatial uses. Cognitive linguistic analyses of polysemous words consider spatial and non-spatial uses to be distinct from each other, if perhaps related by some cognitive principles (Tyler and Evans, 2001). If some but not all participants create these mixed groups, that would suggest that individuals differ in how meaningful they find the spatial/non-spatial distinction to be. Finally, in the case of the data for *above* and *below*, I also assess whether the data indicate whether participants judge TEXT USES of these words to have a spatial or non-spatial (e.g., temporal) meaning.

Quantitative analysis

3.6.1 Did participants agree about how the sentences should be sorted?

Table 9 Pairwise agreement values calculated using Morey and Agresti's adjusted Rand

Target word	Mean	Min.	Max.	Range
Over	0.39	0.22	0.66	0.44
Under	0.65	0.35	0.84	0.49
Above	0.31	0.13	0.55	0.42
Below	0.37	0.01	0.90	0.89

The agreement values shown in Table 9 reveal two interesting outcomes. First, if we examine the 'Mean' column, there is a range of agreement values across the four tasks: some minor, and some major. There is minor variation in mean agreement values across the tasks for *over, above*, and *below*: all hover between 0.3 and 0.4. However, the mean agreement value for the *under* task is substantially higher than the other three tasks, with participants reaching a mean agreement value of 0.65. Second, if we examine the 'Min.', 'Max.' and 'Range columns, we can observe that the extent to which pairs of participants agree with each other is highly variant. For example, as highlighted in the red cell, one pair of participants completing the *below* task reached a level of agreement barely above chance, at 0.01. At the other end of the scale, as shown in the cell highlighted in green, one pair of participants who also completed the *below* task reached a very high level of agreement, achieving an agreement value of 0.90. While ranges in agreement values were observed in the other three tasks, they are particularly great in the *below* task. The possibility that the large sample size in the *below* task resulted in a larger range of agreement values was tested by splitting the sample into three random groups. Table 10 below confirms that this is not the case; the range of agreement values within each sub-group is similar to those observed in the groups who completed the tasks for *over, under* and *above*.

Table 10 Pairwise agreement values for subgroups of participants who completed the *below* task, calculated using Morey and Agresti's adjusted Rand

***Below* sub-group**	**Mean**	**Min.**	**Max.**	**Range**
1	0.38	0.23	0.70	0.46
2	0.56	0.30	0.81	0.51
3	0.19	0.01	0.56	0.56

As noted in section 2.6.5 the lowest acceptable level of inter-participant agreement is the subject of ongoing debate. In the case of WSD research proper, in which human annotations will be used to train a computer, a high degree of inter-annotator agreement is desirable. If we accept 0.8 as the lowest acceptable level, after Neuendorf (2002), the 'Mean' column in Table 9 show that, on average, the extent to which participants agree with each other is unacceptably low. While the 'Max.' columns show pockets of good agreement in the tasks for *under* and *below*, on the whole agreement is really rather poor.

Given that the aim of the present research is to investigate the psychological status of word senses, as opposed to isolating just what the senses of particular words are, the matter of a minimum level of acceptable agreement is somewhat moot. Instead, the scope of the project means that an open analysis of the data is possible, one which seeks to examine what happens when groups of individuals sort the stimuli into groups. However, agreement values remain interesting and useful sources of data, since they reveal variation in the extent of inter-participant agreement. Specifically, low agreement values might indicate individual differences in word senses.

We can perhaps anticipate some inter-participant disagreement in a sorting task such as this. Disagreement has already been observed in WSD research, as noted earlier, and in Experiment 1. However, while some disagreement is certainly expected, the extent of disagreement observed amongst some pairs of participants pushes that expectation to the limit. In the case of some pairs of participants sorting examples of *below*, agreement is barely better than chance. Second, the range of agreement values noted across participants and across the four words suggests that agreement is highly variant. They complement the data gathered in a closed sort task by

Passonneau and her colleagues (2012, 2010) which found that agreement varied within, as well as across, parts of speech.

At this stage, it is unclear what is behind the differences in agreement across the target words. One potential explanation is that the task was too demanding. Asking participants to carefully read and disambiguate 100 uses of a particular word while simultaneously keeping track of the categories they create is a very demanding task. It might be that the scale of the task caused fatigue, boredom, forgetting, semantic satiation, or a combination of these.

3.6.2 How many sense groups did each participant create?

Table 11 Number of sense groups created for each word

Target word	Mean	Min.	Max.
Over	17	12	20
Under	6	3	11
Above	7	2	9
Below	6	2	12

As this table shows, participants varied in the number of sense groups they created; the 'Mean' column reveals variation across the four words, while the 'Min.' and 'Max.' columns show variation across participants. Because this was an exploratory study, I did not aim to balance the range of senses (as I perceived them) represented across the four words, nor did I choose to balance the number of stimuli that exemplified each sense. Removing these kinds of structures should mean that participants' responses are as authentic as possible, and influenced as little as possible by my own intuitions.

The differences in the number of groups created provide some explanation for varying levels of pairwise agreement. Depending on the exact composition of two participants' groups, the creation of more groups by one participant than the other might by default result in lower agreement than if they created the same number. Take, for example, a pair of participants who created five groups which had identical composition, and in addition participant A created one with a set of six sentences, which were split into two groups by participant B. The pair would achieve lower

agreement than if they were to use the same number of groups. Of course, this seems like an unlikely scenario, and it is more likely that participants will create groups that have some, but not complete, overlap. This will mean that differences in the number of sense groups that they create may not have the same effect on pairwise agreement scores across all pairs of participants.

The difference in the number of groups hints at participants' varying tendencies to seek out broad, generic senses amongst the stimulus sentences. A small number of groups created by a participant might therefore suggest that, when sorting the sentences, they look for uses of *below*, for example, that broadly share a meaning. This is not a watertight approach to the analysis of this data, though; as we will see in the following close analyses, participants who use a small number of groups do not necessarily *only* make broad distinctions.

Qualitative analysis

At this point, I direct your attention to the similarity matrices in Appendix 4.[4]

3.6.3 Over

3.6.3.1 Did any meaningful senses emerge in the similarity matrix?

The table below lists the groups that emerge from the similarity matrix, and offers a suggestion – based on my interpretation of the underlying meaning of the target word in each constituent sentence – of the meaning of the sense used in each group. For ease of reference, an example sentence for each group is also provided. The following discussion will consider first the more clear-cut, easily-interpretable groups, before moving onto the groups that are trickier to understand.

[4] Due to the size of these files, they are best viewed on a computer screen.

Table 12 Groups detected in similarity matrix for *over*

Colour on similarity matrix	Group number	Tentative definition	Example sentence
	1	EXCHANGE	James Roberts [q.v.], whose printing business he soon took OVER.
	2	CONTRASTIVE POSITION	Can we have it OVER here mum?
	3	UNCLEAR	The plane flew OVER the city
	4	ARC	Go up that one and jump OVER the other one
	5	COVERING	She pulled the covers OVER herself
	6	FLIP	Turn that steak OVER, it's burning!
	7	FALL	Why did you do that you nearly knocked me OVER then
	8	DURING	We're going to go camping OVER the Easter holiday
	9	MORE THAN	Council house rent arrears amounted to OVER £1m, though they are at long last being reduced.
	10	CONCERNING	Afternoon session there was an equally irritable wrangle OVER a proposal to adopt a law on compliance with the
	11	REPETITION	I don't want to have this fight all OVER again
	12	COMPLETION	I can't believe the weekend is OVER already!

Analysis of the similarity matrix reveals two interesting outcomes. On the one hand, the very light shading seen in some parts of the similarity matrix reveals some overlap in how sentences were sorted, indicating that participants differ from each other in what stimuli they consider to use *over* in an equivalent way, and what stimuli they consider to use *over* in a different way. This disagreement is reflected also in the pale shading within each triangular cluster of sentences. This latter observation indicates that the exact composition of the groups I shall propose here is rather uncertain, meaning that the senses I posit in this chapter are tentative. With that said, the similarity matrix does in some places reveal some groups that have a reasonable degree of agreement and coherence. Further, as demonstrated by the large areas of white space, occurring where no participant assigned a particular pair of sentences to the same category, some sentences are considered highly distinct from certain others.

3.6.3.1.1 Clear cases

The first group, in the top right-hand corner, might be paraphrased as describing an exchange, or a change of hands. Its composition is not wholly agreed upon; for example, *handing over its presidency in December next year* and *He said that's half*

the reason that Brian Tolbrook took over at Tettron ain't it are grouped together only 50% of the time, while the latter sentence and *James Roberts [q.v.], whose printing business he soon took OVER*, are grouped together 100% of the time. On the whole, though, these three sentences are grouped together more frequently than with other stimuli, indicating that they are a distinct group.

The second group, and one which is much larger, contains sentences that, on the whole, describe some kind of contrastive position. It has some pockets of very good agreement, but also some patches of weaker agreement. While most of the sentences describe contrastive position (e.g., *Can you call her over here for a minute?*, and *We got a telephone over here*), some seem less exemplary of this overarching schema; for example, *We often walk over the fields* seems, to my mind, to convey an "across" sense of *over* – of course, however, the findings presented in Chapter 2 demonstrated that my intuitions do not always neatly correspond to those of other speakers. This sentence, along with *I ran over the bridge,* are grouped with the sentences describing contrastive position perhaps by virtue of the fact that similarity matrices are constructed on the basis of pairwise grouping: these two examples are grouped more with *Can you just run it over the road* than with other sentences outside of this second group, which itself is grouped with the other sentences in the group more than sentences outside of the group. Their poor exemplariness manifests in the similarity matrix as low agreement values with most other stimuli in the group.

The sixth group, another group capturing spatial relationships, can be best characterised as representing a FLIP sense of *over*, and is the subject of a good degree of internal agreement.

The seventh group, also spatial in nature, represents a FALL sense of *over*. While it is a fairly discrete group in that sentences that are outside of the group are rarely matched with any of the component sentences, it is not a group that has particularly good levels of agreement. These facts suggest that while participants tend to agree that the sentences are dissimilar to sentences outside of the group, they do not fully agree about how similar the member sentences are to each other.

The ninth group, which captures a non-spatial sense of *over*, comprises sentences using *over* in a MORE THAN sense. Given the conventionality of using *over* to describe positions on numerical scales, it is surprising that this group does not have a particularly strong composition; certainly, there are pockets of excellent agreement, but the domain-specificity of this group encourages us to predict that participants would reach much better agreement about how its members should be sorted. This prediction is not realised, though.

A much stronger group is the tenth cluster, which captures a CONCERNING sense of *over*. Its size – it comprises just four sentences – might explain its high level of agreement. However, in the face of the eleventh group (representing a REPETITION sense of *over*), which comprises just two members but has weaker agreement, it seems that we cannot chalk the strength of this group up to its diminutive size.

The final group, representing a COMPLETION sense, is another group that has patchy agreement. On the one hand, *They think it's all over… it is now!* and *I can't believe the weekend is over already!* have a pairwise agreement of 100%, as do the pair *I am so over him* and *It took him ages to get over the flu.* On the other, the pairwise agreement amongst these four sentences is much weaker. The reason for this might be that while all capture a sense of completion, they might be divided more finely to specify a COMPLETION + RECOVERY sense, represented by the latter two sentences. In the case of the former pair, a simple COMPLETION sense is invoked.

3.6.3.1.2 Trickier cases

The composition of the third group is rather mixed, making any underlying meaning unclear. It includes examples which seem to use *over* in varying ways; for example, *15 OVER 3 equals…?*, *The family portrait sits OVER the* fireplace and *The plane flew OVER the city*. While within the middle of the triangular cluster there are darker patches, indicating a higher degree of agreement, the group as a whole seems rather uncertain. Indeed, some members are grouped together at a level lower than we can expect by chance.

The fourth group is also somewhat mixed, with rather a lot of overlap with the fifth group. However, the group tends to capture an ARC sense of *over*, with sentences

describing atypical arcs (e.g., *You almost ran over that rabbit!*), partial arcs (*Fall over the bridge*) and full arcs (*The quick brown fox jumped over the lazy dog*). Collectively, these arcs are captured by the schematic illustration in Figure 33. In this figure, the dashed line indicates the optional aspect of the arc trajectory; for example, as in the partial arc described by *Fall over the bridge*.

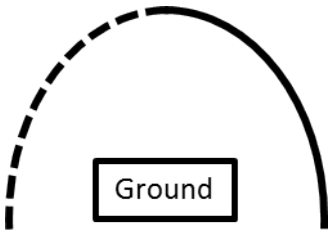


Figure 33 Illustration of the underlying spatial configuration captured by group 4, ARC.

As noted, there is some overlap between the members of the fourth and fifth groups, the latter of which might be best described as representing a COVERING sense of *over*. An explanation for the overlap in membership of the fourth and fifth groups might be that the spatial configurations sentences in each group describe share a basic configuration: the extension of a figure over a ground. In the case of most of the sentences in the fourth group, that extension is specified further to describe an arc-shaped *movement*, as illustrated in Figure 33. In the fifth group, the extension of the figure across the body of the ground is *static* and widespread, as illustrated in Figure 34.

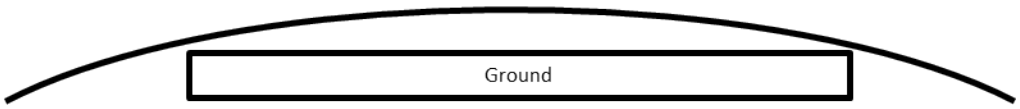


Figure 34 Illustration of the underlying spatial configuration captured by sense group 5, COVERING

There is mixed agreement over the composition of group eight. At the far end of the group it appears that the grouping of some sentences has a very good degree of agreement. For the most part, the sentences in the group use a DURING sense of *over,* for example *Let me think about it over the course of the day*. The sentence, *Can you look over this report for me?* (which I'll describe for the moment as an example of the THROUGH sense of *over*) is rather distinct from the other sentences, but it is

perhaps less distinct than the sentence *I mean that's Suffolk all over isn't it, really for you?*. Both DURING and THROUGH senses capture a sense of motion through some domain; in the case of the sentences in this group, through the temporal domain, and through the domain of a body of text. They share common features: a start and end point, and an intermediate zone that is canonically navigated in a linear fashion. The figures below illustrate this commonality of meaning.

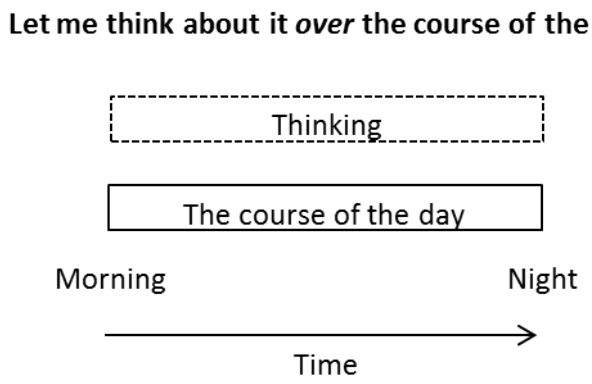


Figure 35 Schematic illustration representing *Let me think about it over the course of the day*

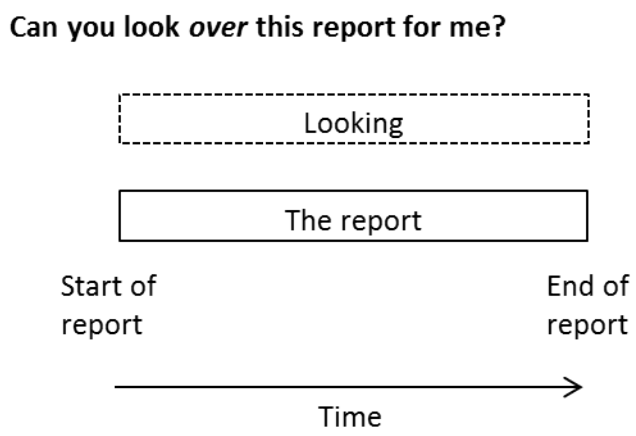


Figure 36 Schematic illustration representing *Can you look over this report for me?*

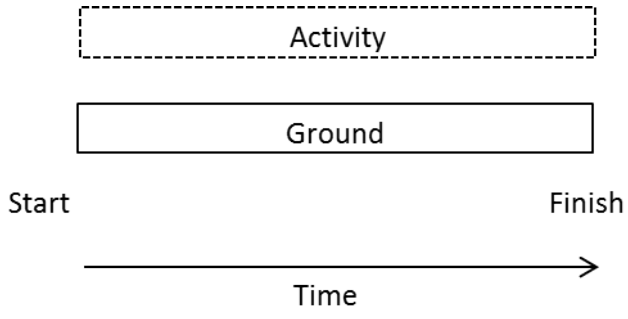


Figure 37 Schematic illustration representing examples in Figure 35 and Figure 36

If we overlook the *Suffolk* sentence, perhaps we can rationalise the composition of this group. This sentence seems so at odds with the rest of the group that we may be better off considering this as an outsider; the lone member of an exceptionally small group that the similarity matrix, perhaps for technical reasons, cannot represent.

3.6.3.2 Lumping and splitting

Arising from criticism surrounding excessively fine-grained sense distinctions proposed in cognitive linguistic literature, a contentious issue within the literature on polysemy is that one person's notions of what constitutes a distinct sense may be rather different from that of another person. In other words, there may be variations in number of senses a group of speakers find meaningful. So while one person may identify a large set of minutely distinct senses in a set of, say one hundred sentences, another may find five to be meaningfully distinct. The people occupying these polar positions are described as "splitters" and "lumpers", respectively (e.g., Hanks, 2000). We can use the following sentences to illustrate these definitions:

6. I can't believe the weekend is *over* already!
7. It took him ages to get *over* the flu.

The decisions made by some of the participants in this study suggest that these sentences may be grouped separately, or together. Together they relate to the ceasing of some event: in example 6 it is a temporal event (the weekend), in example 7 it is a condition event (the flu). Example 6 can be further specified, though, to account for a sense of recovery that accompanies the ceasing of the event. It is a distinction that some, but not all, participants find sufficiently meaningful to license the creation of separate groups.

Some participants demonstrated tendencies for both lumping *and* splitting, rather than adopting a single sorting strategy. For example, participant O6 separates examples of *over* that describe some form of temporality into three groups, labelled 'time', 'recover' and 'through'[5]. The 'time' group is interesting in that it conflates what I consider to be examples of three distinct senses: COMPLETION (e.g*., I can't believe the weekend is over already!*), DURING (e.g., *We're going to go camping over*

[5] Inverted commas are used to indicate the sense labels created by participants

the Easter holiday) and REPETITION (e.g., *I don't want to have this fight all over again*). The 'recover' group captures the distinction identified in examples 6 and 7 above. 'Through' comprises only one sentence, *Let me think about it over the course of the day*. It is unclear what makes this final sentence distinct from the sentences in the 'time' group. Beyond this subset, we can observe some lumping: when sorting spatial examples of *over*, the logic behind their decisions is not always clear. The fact that I find it difficult to follow the logic behind this participant's sorting adds weight to the claim made in Chapter 2 that participants and I differ in what examples of *over* we consider to have equivalent meanings.

The presence of fine distinctions contributes to the debate in cognitive linguistic literature over whether some linguists' predilection for splitting is shared by naïve speakers (Sandra and Rice, 1995). Work by Cuyckens, Sandra, and Rice (1997) and Sandra and Rice (1995) considering the psychological reality of polysemy accounts made in cognitive linguistics has begun to shed light on naïve speakers' tendencies to lump and split examples of polysemous words. In both sets of studies, the authors found evidence that participants did make some fine distinctions. The fact that individual participants – here and elsewhere in this study – use both splitting *and* lumping strategies adds detail to this issue: it is not the case that linguists split and participants lump; a single participant may split where a linguist would lump, and vice versa.

While there is a tendency for participants to create large and seemingly undifferentiated groups of spatial sentences, not all participants adopted this approach. Participant O9 was unusual in their decision to split spatial uses on a very fine basis: they separated examples describing static versus dynamic relations, the presence of direction, and whether a fall was implicated. These are not distinctions that I would make, which provides further evidence in support of the conclusion that at least some of the senses of these words that I find to be meaningful are not shared by other speakers. There is clearly growing support for this interpretation, adding weight to the conclusions drawn in Chapter 2.

3.6.3.3 Did people distinguish between spatial and metaphorical uses?

Mixed sense groups – that is, groups which to my mind captured examples of *over* that describe both spatial scenes and non-spatial scenarios – were found in the groups created by all except one (O9) participants. The tendency for participants to create mixed groups suggests that the distinction between concrete and abstract scenarios is not necessarily the most meaningful one when deciding how a group of sentences should be sorted. Indeed, most of these participants (O1 is the exception) created *multiple* mixed groups – as many as three by some participants. The presence of multiple mixed groups rules out the possible explanation that spatial and non-spatial sentences were grouped together accidentally and instead supports the interpretation that whether or not they describe spatial configurations is not the ultimate deciding factor when distinguishing between meanings. In the context of this study, it therefore seems that the meaningfulness of the difference between spatial and non-spatial uses varies between individuals.

This idea seems rather controversial, given that the distinction between the concrete (i.e., spatial) and abstract (i.e., non-spatial) is arguably a fundamental one in theoretical treatments of polysemy in cognitive linguistics – and, indeed, in conceptual metaphor theory (Lakoff and Johnson, 1980). If the use of mixed sense groups was anomalous, and restricted to perhaps one participant, then we might overlook that decision as an anomaly. In the face of a general tendency to create at least one mixed sense group, though, this tendency must be acknowledged. The significance of this finding is discussed in more detail in section 3.7.3 below.

3.6.4 Under

3.6.4.1 Did any meaningful senses emerge in the similarity matrix?

The matrix for *under* can be divided into two clear sections: the first, positioned at top-left part of the matrix, is a large, generally undifferentiated group of spatial uses. The second, positioned at the lower-right part, comprises a set of smaller groups of varying degrees of distinctness. The table below lists the groups that emerge from the similarity matrix, and provides a suggestion of what sense is used in each sentence in the groups. For ease of reference, an example sentence for each group is also provided. The following discussion will consider first the more clear-cut, easily-interpretable groups, before moving onto the groups that are trickier to understand.

Table 13 Groups detected in similarity matrix for *under*

Colour on similarity matrix	Group number	Tentative definition	Example sentence
	1	GENERIC SPATIAL CONFIGURATION	She kicked him UNDER the table
	2	LESS THAN	She can only go on the ride if she's UNDER 10
	3	UNDER AUTHORITY/ COMMAND OF	She's got a whole team working her UNDER her now
	4	MIXED METAPHORICAL SENSE	UNDER the terms of this new policy, we should get a refund
	5	INTERNAL FORCE	I was UNDER the impression she'd already quit
	6	EXTERNAL FORCE	Her workload is currently UNDER review

3.6.4.1.1 Clear cases

As noted, the first group is very large, and in fact encompasses all spatial uses of *under*. On the whole, agreement that every sentence in the group should be sorted with every other sentence in the group is very high – for the most part, agreement is 88% or higher. However, pockets of weaker agreement, at or a little over chance level, are observed. These two outcomes indicate that, by and large, most of the spatial sentences in the task are equally good exemplars of a spatial sense of *under*, though some are rather a lot less exemplary, which results in disagreement over where they should be sorted. We may therefore conclude that participants tend to agree that most of the examples of *under* in the stimuli in this group are similar in meaning to each other, but there are cases when participants' judgments over similarity or equivalence diverge.

The second significant group in the matrix displays a high degree of internal consistency, as represented by agreement values of 75% or higher, and a low degree of overlap with other groups. This second group captures a LESS THAN sense of *under*, in which each sentence describes a position on a numerical scale.

The fifth group, comprising just two sentences, is a very discrete group representing an INTERNAL FORCE sense of *under*; in both of these examples, the speakers' actions are shaped by the internal forces of their beliefs.

The final group, and one which has mixed levels of internal agreement and rather a lot of overlap with other groups, can be described as representing a SUBJECT TO sense

of *under*. In each case, an abstract or concrete figure is subject to the external force of an abstract ground. One exception, perhaps, is the sentence *moves are under way to establish a modern Socialist party*. *Under way* is a specific collocation that does not seem to be quite as good an exemplar of a SUBJECT TO sense as other examples in the group. This is reflected in its weak membership: agreement that it should be paired with other members is rarely above chance level.

3.6.4.1.2 Trickier cases

The third group tends towards representing an UNDER THE AUTHORITY/COMMAND OF sense. However, *I booked it under my wife's name* does not quite fit with this definition, and therefore necessitates some caution in interpreting this group. A further issue in this group is that two sentences, *He whispered the password under his breath* and *people have been living on the breadline and under the breadline*, while positioned separately from the group in the similarity matrix, have strongest agreement with the other sentences in the group, compared with the rest of the sentences in the matrix. It is not clear why the algorithm behind the similarity matrix has positioned these stimuli away from the rest of the group.

The fourth group also seems rather mixed, comprising sentences which use ACCORDING TO/COMPELLED BY (e.g., *Under the new regulations, the students had to sign in each week*), IN THE CLASS OF (*I'm going to file that comment under "P" for pathetic*) and a particular CONDITION sense, represented by *I'm under his protection now*. This makes it somewhat hard to argue that this group represents a particular sense.

3.6.4.1.2.1 An alternative interpretation?

The high degree of overlap between the final four groups in the similarity matrix encourages an alternative interpretation: perhaps the matrix comprises three groups: one representing a very generic spatial sense, one a well-defined numerical scale sense, and a final group capturing examples of *under* that describe power structures and their associated effects.

If we were to posit the third large triangle as a single group, we must acknowledge that, within that, there are pockets of extreme proximity and similarity, within which sentences are very distinct from other members in the mega-group. However, there is

a meaningful overlap across most of the sentences. They generally use *under* to describe power structures or the effects thereof. People in positions of authority (such as those in *I'm working under the direction of the area manager*) are able to implement policies such as the one described in *Under the old system, commission payments for transactions*, and they can impose external forces such as those described in *We're under real pressure at the moment*. Perhaps the sentences in the penultimate group, which represent an INTERNAL FORCE sense, are members of this mega-group by virtue of their relationship to the final, EXTERNAL FORCE group. This outcome suggests that the need for fine distinctions in effective communication may not be necessary; instead, broader groups encapsulating smaller sub-groups, may be sufficient. This is compatible with the argument made by Ide and Wilks (2007) concerning the coarse level of meaning humans and computers generally access in order to achieve successful understanding, an idea which is discussed in section 4.1.1.

3.6.4.2 Lumping and splitting

Some degree of both lumping and splitting is observed in most of the participants' sorting decisions. Participant U2 adopts a very fine-grained approach to identifying sense boundaries, particularly within the spatial domain. For example, they distinguish between sentences that use *under* to describe 'covering' relationships, uses describing general vertical configurations, usually without contact, a one-member group labelled 'behind', in which a non-canonical horizontal configuration is described, and a final group labelled 'through', which captures sentences in which a figure passes under and out of a ground such as a bridge. The non-spatial groups do not show such fine-grained distinctions; while they create separate groups for 'control', 'governed by' and 'with', some of their non-spatial groups seem to be more catch-all.

Elsewhere in the group of participants, there is evidence of extreme lumping. For example, participant U5 makes just three distinctions, which capture spatial, non-spatial and mixed sense-type (i.e., a mixture of both spatial and non-spatial) items. This approach means that they are classified almost by default as a lumper. On closer inspection, the spatial and non-spatial groups do indeed reveal what I would consider to be undifferentiated grouping. The creation of a mixed sense group indicates that

the participant feels that there is a meaningful distinction between the members of this group and members of the other two groups, and is not therefore solely concerned with a binary distinction between spatial and non-spatial meanings. Accordingly, it seems unjustified to casually classify the participant as a lumper. It would have been helpful to have spoken with the participant to understand why a third group was created; as it stands, we cannot know the dimension on which this participant splits items.

As might be predicted from analysis of the similarity matrix, close analysis of individual participants' sorting decisions shows that, when participants do allocate non-spatial sentences to a discrete group, this is achieved with most coherence and consistency when the sentences describe positions on numerical scales. Indeed, when participants make more than one distinction amongst non-spatial items, at least one of those distinctions is always made to separate these types of sentences from other non-spatial examples.

3.6.4.3 Did people distinguish between spatial and metaphorical uses?

On the whole, participants generally distinguished between sentences that described spatial configurations and relationships, and those which did not. Across the eight participants, only one created a group that captured both spatial and non-spatial examples. This outcome suggests that the distinction between spatial and non-spatial senses of *under* is a highly salient one, meaningful to most – but not all – participants.

The tendency for participants to create a single, seemingly undifferentiated group of all of the spatial uses of *under* (done so by five of the seven participants) suggests that amongst these participants, there was a sense that as long as the sentence at hand describes a spatial configuration, the use of *under* is equivalent in meaning to all other sentences describing spatial configurations. It is only in the case of participant U2 that we can observe finer distinctions at work: this participant seems to divide the stimulus sentences according to whether a 'covering', generic vertical, 'through' or 'behind' relationship is described.

3.6.5 Above

3.6.5.1 Did any meaningful senses emerge in the similarity matrix?

As was the case in our analysis of the similarity matrix for *under*, we can divide the similarity matrix for *above* into two distinct sections. The first section, which takes up the majority of the top half of the matrix, comprises the majority of the spatial uses of *above*. The second, occupying the lower half of the matrix, is a set of more discrete groups representing spatial and non-spatial senses. The table below lists the groups that emerge from the similarity matrix, and provides a suggestion of what sense is used in each sentence in the groups. For ease of reference, an example sentence for each group is also provided. The following discussion will consider first the more clear-cut, easily-interpretable groups, before moving onto the groups that are trickier to understand.

Table 14 Groups detected in similarity matrix for *above*

Colour on similarity matrix	Group number	Tentative definition	Example sentence
	1	LOUDER THAN	It's really hard to hear you ABOVE the music
	2	GENERAL SUPERIOR SPATIAL POSITION	We never take from ABOVE the brow, just below
	3	TEXT USE	In summary, the detailed empirical word outlined ABOVE has enabled another general law
	4	NUMERICAL SCALE	Train fares have risen ABOVE inflation
	5	ABOVE ALL	Then ABOVE all we do need a very large membership
	6	RANK/HIERARCHY	The orders came from ABOVE

3.6.5.1.1 Clear cases

The first group, which has an excellent degree of agreement and coherence, comprises sentences that use a sense of *above* to describe a figure sound that is (e.g., *then, above the cracking of the fire, he did hear something*), or is required to be (e.g., *It's really hard to hear you above the music*), louder than a ground sound. While some participants matched members of this group with sentences from other groups, this was done rarely.

The sense captured by the third group in the matrix referred to in this thesis as a TEXT USE sense of *above* (these uses will be discussed further in section 3.6.5.4). Its internal consistency and relative distinctiveness from other groups might be explained on the basis that participants will likely recognise the usages as belonging to the same modality, specifically, written text. It is also a sense which I would

anticipate that all readers of English encounter frequently, given that it is used across all forms of written media, from leaflets and forms to newspapers and books. Further, it has a very specific purpose: to direct the reader's attention – be that mental or visual – back to a part of a preceding portion of the document. Perhaps these characteristics collectively allow participants to recognise the equivalence in meaning amongst the members of the group, and their distinctiveness from other sentences.

The fourth group is varied in the extent to which participants agree that its members should be grouped together, but this does not prevent a sense emerging from the cluster. Each of the sentences describes a position on a numerical scale. One exception might be the sentence *I know I'm pushing above my weight*; this is a conventionalised collocation, but one which arguably takes its roots in the description of a numerical scale.

The penultimate group in the similarity matrix enjoys a very high degree of agreement, which is likely the result of the fact that *above* is used in its member sentences as part of a particular collocation: *above all*.

3.6.5.1.2 Trickier cases

It is unclear whether what I propose to be the second group is one group, or comprises a number of smaller groups. While there are certainly sets of sentences that arguably match each other better than with other sentences in the group, the group as a whole has so much internal agreement and overlap that making finer distinctions within the group may be an unnecessary and perhaps artificial interpretation. Instead, I propose that the sentences in this group reflect a generic and underspecified spatial sense of *above* in which a figure is located in a position superior to a ground. Note that this configuration need not be strictly vertical; as *mysterious Glastonbury Tor, which towers above the flat landscape of the Somerset Levels* demonstrates, as long as the figure is in a raised position, the configuration can be diagonal or vertical.

The group of three sentences immediately below group 3 are sorted together with a good degree of agreement. Its position in the matrix, separate from other groups,

may encourage us to conclude that it does indeed represent a distinct sense. However, if we shift our view to the left of the matrix, we can see that the constituent sentences were sorted with some regularity with members of the large generic spatial group (group 2). The sentences in this group also describe spatial configurations, but what sets them apart is their specificity; the sentences each define the precise dimensions of the spatial scene described. Are these sentences members of a group that is distinct from the large, second group? Given that the items in this smaller group are sorted with some members of group 2 with a good degree of agreement, I would argue not. Certainly, they are considered less similar to some members of the large second group than others, as indicated by the paler shading than the dark cells that characterise much of group two. We might consider these to be "fringe" members of the group, and members by virtue of their similarity to some – but not all – members of the group.

The final group arising from the sorting task shows patches of high agreement, but overall some disagreement over how similar certain members of the group are to certain other members. The result is a group with varying levels of pairwise agreement. There is evidence here that while I may take a splitting approach when categorising these sentences – I pick out distinctions between examples depicting a personal sense of superiority (exemplified by *She likes to think she's above all that silly gossip*), a more objective sense of qualitative superiority (*provide working conditions which are or above the average of its locality*), and rank (*I used to be his boss, but he works above me now*) – these distinctions do not seem to be meaningful to the participants in this task. Instead, the groups in the matrix suggest that on the whole, participants find these examples to be similar in meaning, though not all equally similar. While I find the distinctions specified above meaningful, I can equally understand why these distinctions may be overlooked in favour of creating a single group. A person or object that is positioned above another in an organisational hierarchy or ranking table is expected to be considered superior to those persons or objects positioned lower. Ranks and hierarchies and superiority (self-regarded or otherwise) are therefore tied together, making a distinction perhaps less meaningful to some participants.

3.6.5.2 Lumping and splitting

As we have seen in examinations of other words, participants in this task cannot always be described in terms of lumpers or splitters; some participants tend towards lumping sentences into large, seemingly undifferentiated groups, while others sort their sentences in ways that suggest both lumping *and* splitting tendencies. Participant A8 is a good example of a participant who demonstrates both tendencies. They created six groups for the sentences, comprising one spatial group, one non-spatial group, and four mixed groups. The spatial group consists of sentences which almost exclusively describe two-dimensional relationships (e.g., *The raven perched above the door*). This factor of dimensionality is clearly an important one to this participant. However, other distinctions, such as the distance between figure and ground, or the degree of verticality in their relationship, are not of issue. Participant A2 adopts both approaches also; they create groups reflecting fine distinctions in non-spatial uses, to distinguish between relative loudness, importance, quality, and superiority/rank. In contrast, their spatial sentences are divided only according to whether they describe configurations with or without contact, and even this distinction seems to have been imperfectly executed. In contrast, participant A4 adopts a lumping approach. They create two groups, labelled as 'physical' and 'metaphysical', which would suggest that the participant is conscious of the distinction between the spatial and the abstract. However, this binary decision is imperfectly executed; one group consists of an undifferentiated collection of non-spatial items, while the other is a mixed group of spatial and non-spatial items.

3.6.5.3 Did people distinguish between spatial and metaphorical uses?

Mixed-sense groups were observed in the groups created by six of the eight participants. While most participants did use mixed-sense groups, they were relatively unusual occurrences; most participants created just one (one [A8] created four, and did not compensate for this heavy use of mixed-sense groups by creating more unmixed groups than other participants). One participant, A4, did not create any groups that exclusively captured spatial examples. Instead, the spatial items were captured within a group combining spatial, non-spatial and text use examples. This decision was unusual, though, and all other participants created at least one group to capture spatial examples, and another to capture non-spatial examples.

The presence of these mixed-sense groups in addition to unmixed groups of both spatial and non-spatial sentences therefore suggests that while participants are sensitive to the distinction between uses of *above* to describe spatial configurations and non-spatial phenomena, that distinction is not always the most important factor when deciding whether a set of examples of *above* are equivalent in meaning in all participants.

3.6.5.4 Were TEXT USES classified as examples of a spatial or non-spatial sense?

Before this question is answered fully, it is necessary to explain how these uses were analysed. When analysing the sorting decisions each participant made, TEXT USES of *above* were checked to establish whether the labels participants gave the groups described the underlying meaning in spatial terms (for example, 'underneath', or 'further up') or in another way (for example, 'later in a document', which has a temporal meaning). They were also studied to note whether they were sorted with spatial or non-spatial items. These observations shaped the way I will talk about how these particular uses were classified in the mind of the participant. When TEXT USES were sorted into a group with spatial items, or labelled in spatial terms, I classified them as representing a spatial sense group. Where they were sorted into a group with non-spatial items, or labelled in non-spatial terms, they were classified as a non-spatial sense group.

Close inspection of the labels participants used reveals interesting insights into what meaning *above* has in these uses in the mind of the participant. Participant A3 grouped these sentences together under the label 'below previous sentences'. In this label, the problem of establishing just what the nature of these uses are is captured: do they describe physical, spatial configurations, or abstract, temporal relationships? A3's label indicates that, in this participant's view, these uses reflect temporal *and* spatial configurations. I would agree with A3's description of this group: this line of text is positioned in a lower location than, for example, the end of the last paragraph, but it is also encountered at a later time than the reader encounters that higher section of text. It is, therefore, a sense that is both spatial and temporal in nature.

If A3 observed that these TEXT USES capture both the temporal and the spatial, it is not an observation that is evident in the responses made by other participants. By and

large, participants positioned these sentences in the domain of the non-spatial; three participants sorted them all using a temporal label, and two sorted them with other, non-spatial sentences. Only one participant, A7, sorted all of them with non-spatial items. The final participant, A8, split them into four groups and mixed them with both spatial and non-spatial items.

The sorting and labelling decisions indicate that most participants in this task view at least some of the TEXT USES of *above* as capturing a temporal meaning. While their responses do not preclude the participants' recognition of these examples as also capturing some spatial meaning, it is arguable that what sets them apart from other sentences in the task is that they describe a temporal relationship, and that if the participants recognise that a spatial configuration – i.e., the relative position of the referring and referent text – underlies this meaning, the fundamental spatiality of the example sentences is less salient, or is of less pertinence, than the temporal configuration the sentences describe.

3.6.6 Below

3.6.6.1 Did any meaningful senses emerge in the similarity matrix?

The similarity matrix produced for the task reveals eight senses of varying degrees of distinctness. The table below lists the groups that emerge from the similarity matrix, and provides a suggestion of what sense is used in each sentence in the groups. For ease of reference, an example sentence for each group is also provided.

Table 15 Groups detected in similarity matrix for *below*

Colour on similarity matrix	Group number	Tentative definition	Example sentence
	1	QUANTITATIVELY LESS THAN	The loss is a little BELOW £3,200.
	2	QUALITATIVELY LESS THAN	They were of BELOW average ability.
	3	RANK	BELOW the rank of sergeant, the pay is terrible.
	4	TEXT USE	Fill in BELOW all the tasks that you do in a typical day.
	5	VANTAGE (3D)	She looked out at the city BELOW like a living map.
	6	LAYER (3D)	They opened a shop BELOW the Redcar office.
	7	LOWER THAN (2D)	There was a large scratch BELOW the driver's window.
	8	GEOGRAPHICAL	Anything BELOW the Watford Gap is the south.

Unlike the preceding analyses of *under, over* and *above*, the groups that emerged in the task for *below* were all fairly uniform in their intelligibility. For that reason, only a short outline of the groups is presented here.

The first group, at the top-left of the matrix, captures a QUANTITATIVELY LESS THAN sense of *below*, covering numerical values of any kind. The second represents a QUALITATIVELY LESS THAN sense, capturing sentences describing inferior positions on a qualitative scale. The third group captures examples that depict ranks in organisational hierarchies and ranking more generally. In combination, labelling this sense as RANK seems appropriate. The fourth group combines all sentences that use *below* to refer to the position of a piece of text.

After this group begin the spatial senses proper; the fifth group can be best described as capturing sentences describing a VANTAGE-like perspective on three dimensions, in which the figure and ground are configured in a more diagonal arrangement. The sentences in the sixth group collectively describe LAYER-like configurations, while the seventh group captures two-dimensional spatial relationships best described by the term LOWER THAN. The final group captures a particular type of use of *below*, which is invoked to describe the position of a geographical figure in relation to a geographical ground.

While inspection of the similarity matrix allows these groups to be identified, they are not equally identifiable. In other words, some groups are more clearly distinguished from their neighbours than others. There is a notable distinction in the clarity of the groups that emerge from the non-spatial versus spatial sentences. Non-spatial sentences cluster into more readily-defined groups than the spatial sentences, as can be seen in the light shading underneath the darker triangles that form below the non-spatial sentences. The shading of the triangle under the spatial sentences is much more inconsistent. It is only when one "zooms in" and looks at the values within the larger triangle that smaller triangles, representing more defined senses, emerge.

3.6.6.2 Lumping and splitting

Participants demonstrated tendencies for both lumping and splitting. For example, participant B1 created just five groups to capture the sentences presented. The non-spatial sentences were split in two: the majority were in one undifferentiated group, indicating a tendency for lumping, but one sentence stood alone in a distinct group, indicating a tendency for splitting. This tendency is repeated in their sorting of spatial sentences: one group contains sentences describing all two-dimensional configurations, and the other contains all three-dimensional configurations. These are not distinctions that all participants find meaningful. However, the finer distinctions others found sufficiently meaningful to license more precise groups were apparently not meaningful to B1. We have, then, in this one participant a tendency to "lump" examples of *below* together without compromising acknowledgment of distinctions that they find meaningful. If the participant was a true lumper, it seems unlikely that the single-member non-spatial group would have been formed. Equally, the participant has taken care to attend to the dimensionality of the spatial scenes described in the stimulus sentences, and used this as a guide when deciding which sentences should be grouped together. Any other distinctions – distance between figure and ground, the nature of the figure and ground, for instance – are overlooked and are not considered sufficiently meaningful to license further groups. This is not an isolated case; other participants demonstrated that they were mindful of some fine distinctions, but not others.

3.6.6.3 Did people distinguish between spatial and metaphorical uses?

Amongst the 21 participants, mixed-sense groups were rare: only seven were created, and two of these were created by one participant (B20). In this way, *below* is therefore similar to *under*; participants who sorted examples of *under* created just one mixed-sense group between them. For the majority of participants in this task, then, it seems that the distinction between spatial and non-spatial uses of *below* was an important marker of difference, and thus acted as the fundamental factor when deciding whether or not sentences should be grouped together or split apart. However, as participant B20 in this task shows, and as has emerged throughout this chapter, the distinction between spatial and non-spatial meaning is not *necessarily* meaningful enough to divide the stimuli into finer groups.

Close inspection of the sorting decisions made is a time-consuming task, but one which provides interesting insights. One particularly interesting decision made by participant B6 is their grouping of sentences that typically describe a change of position, whether that is abstract or physical. For example, they grouped together *The value fell below that of the original investment* and *The crocodile sank below the surface*. We can picture both scenarios in terms of levels: the level of the *original investment* on a numerical scale, and the level of the *surface* of the water, measured on a depth scale.

3.6.6.4 Were TEXT USES classified as examples of a spatial or non-spatial sense?

In the preceding discussion of the results for *above*, it was observed that participants tended to describe TEXT USES of *above* in non-spatial – specifically, temporal – terms, or sort them only with sentences that used a non-spatial sense of *above*. Only one participant gave TEXT USES of *above* a spatial label. Given the relationship between TEXT USES of *above* and *below*, which differ only in that they are opposites, it seems reasonable to expect the same patterns to emerge in an analysis of how TEXT USES of *below* were sorted and labelled.

The opposite pattern emerges when we look at how TEXT USES of *below* were sorted. No participant sorted TEXT USES only with non-spatial items, nor did any participant label these items in non-spatial terms. Instead, where TEXT USES were not considered a distinct group they were labelled in spatial terms, or sorted with spatial items. Accordingly, while these results do not rule out the possibility that participants may

also recognise the temporal nature of these text markers, taken at face value it seems that the spatial nature of these examples of *below* is more meaningful to these participants. This finding is consistent with the observation made in section 2.7.2.5. The significance of this contrasting outcome is considered in section 3.7.4 below.

3.7 General discussion

The aim of the study reported in this chapter was to investigate the possibility that there are individual differences in word senses, which was posited as a possible explanation of the results presented in Chapter 2. In that chapter, it was noted that participants did not reliably agree with me about how examples of *over, under, above* and *below* should be categorised when the categorisation criterion is the meaning of the word in context. It was speculated that this might be due to the fact that the senses I find meaningful differ from those of other native speakers, but that native speakers share a set of senses of these words. An alternative interpretation, that disagreement with my sense distinctions might be due to individual differences in word senses, was also proposed. This interpretation was supported by the fact that participants and I varied in the extent to which we (dis)agreed, suggesting that some participants and I may have at least some senses in common, while other participants and I might differ more extensively over what the senses of a given polysemous word are. In the face of this possible explanation, this study explicitly tested whether or not participants may have different word senses. This aim was achieved recruiting native speakers of English to complete an open sentence-sorting task. This type of task reveals which sentences participants judge to use the same sense of the target word, and which they judge to use different senses of that word.

Statistical analyses carried out using Morey and Agresti's adjusted Rand revealed variation in how well pairs of participants agreed about how the stimuli should be sorted. In the case of *under* and *below*, there were cases in which certain pairs of participants reached a level of agreement in excess of the 0.8 minimum Neuendorf (2002) recommends as reflecting an acceptable level of agreement. On the whole, however, mean agreement scores fell below that; in the case of *over, above* and *below*, mean agreement did not exceed 0.39. On the whole, the study revealed varying levels of agreement both across pairs of participants, and across the four words studied. Good agreement is taken to suggest that word senses are shared, and

poor agreement is taken to suggest that people assign different meanings to the stimuli, in turn suggesting that the senses they find meaningful differ. The data therefore suggest that, on the whole, there is considerable disagreement about what the meaning of the polysemous word in the context of each sentence is, indicating individual differences in word senses.

Qualitative analyses were also carried out, addressing whether any clear senses could be identified using similarity matrices capturing all participants' sorting decisions; whether or not there are material differences in the level of granularity participants apply when categorising the sentences; whether or not participants agree that there is a meaningful distinction between spatial and non-spatial uses of a polysemous word; and the semantic status of TEXT USES of *above* and *below*. These qualitative analyses are brought together in the sections below.

3.7.1 Agreement about the senses of *over, above, under* and *below*

Disagreement about how examples of *over, under, above* and *below* should be sorted suggest two key insights about the nature of word senses. First, the data indicate that participants do not agree what the senses of these words are. Qualitative support for this interpretation is presented in sections 3.6.4 to 3.6.6. If they did, we would expect much greater consistency in their sorting decisions, which would be reflected in clearer similarity matrices, more consistency in the number of groups created, and higher agreement values. Second, as reflected in Table 11, which shows differences in the number of groups each participant created, individuals appear to differ in the number of distinctions they make. Related to this finding is the observation that some participants attend more closely to minute variations in examples of these words than others do.

This finding complicates the position of established analyses of these words, for example those by Tyler and Evans (2001) for *over* and Evans and Tyler (2005) for *over, under, above* and *below*. These studies, as well as other analyses of polysemous words that rely on the author's intuitions, principled as they may claim them to be, seem incompatible with the finding that not only do participants fail to reliably agree with me, in my capacity as a linguist, about what the senses of these words are, but they also disagree with each other. In the case of Tyler and Evans'

studies of these words, it seems that while the distinctions they propose seem logical and are distinguished using articulated decision principles, the distinctions they posit are unlikely to reliably converge with what non-linguists recognise to be meaningful.

3.7.2 Lumping and splitting

As shown in Table 11, there is considerable variation both across participants, and across the four words studied, about how many groups the sentences should be divided into. At this point, it is unclear what is behind participants' differing judgments of the number of sense groups and agreement about how sentences should be grouped together. Ide and Wilks (2007) have proposed that finely-grained senses are only accessed when absolutely necessary for understanding. Unless a fine-grained sense is needed for accurate disambiguation, a more coarsely-grained sense will suffice. The data shown in Table 11 above do not necessarily challenge this proposal, but do complicate it somewhat. If a participant uses few groups, this might indicate that the few distinctions that are used are all that are necessary to disambiguate successfully. A decision to create more groups may reflect participants' belief that there are finer-grained distinctions that make meaningful delineations between uses. In terms of Ide and Wilks' proposal, the number of distinctions participants ultimately feel the need to make in order for disambiguation to be successful therefore differs across individuals.

However, how can we explain the difference in the number of groups created in the *over* task, compared with those created in the tasks for the other three words? Certainly, the range in the number of groups is similar to those observed in the data from the other three tasks, which provides support for the interpretation that individuals vary in the level of granularity they judge necessary to discriminate between senses. The minimum, maximum and mean numbers of groups, on the other hand, are very different. An explanation for this outcome might be that there is more semantic variation in the stimuli in the *over* task – and indeed, in how *over* is used in natural language. Certainly, the Oxford English Dictionary records more senses of *over* than it does the other three words. Perhaps the fine-grained senses of *over* cannot be easily collapsed into fewer coarser groups. The similarity matrix for *over*, when compared with those for *under, above* and *below*, supports this interpretation. As noted in section 3.6.4.1, there is a large amount of white space in the matrix,

which is created when no participant assigns a particular pair of sentences to the same group. This indicates that all of the participants found many possible sentence-pairings to be inappropriate, suggesting that much of the stimuli were considered highly distinct from each other. This coincides with groups with generally rather strong membership. Between these two observations, it seems that groups are generally considered rather distinct from each other. It is only in the case of three groups, UNCLEAR, ARC and COVERING, that there is a more notable degree of overlap in group membership. The same cannot be said for the similarity matrices produced in the other three tasks. In those matrices, there is considerably less white space; in the case of *below*, there is none at all, showing that all sentences were paired together by at least one participant. While there are certainly distinct groups, as discussed in the qualitative analyses earlier in this chapter, there is also a considerable degree of overlap in group membership, which indicates that some groups may be collapsible into broader, coarser groups.

On a somewhat different note, the results of these tasks indicate that notions of "lumpers" and "splitters" is a false dichotomy: individuals may be lumpers, they may be splitters, or they may be both. An empirical study of the polysemous words *in, on* and *at* by Sandra and Rice (1995) indicates that non-linguists *do* make use of some fine sense distinctions. However, it did not establish whether participants tended towards either fine-grained distinctions or broader classifications, or whether participants exhibited both sorting strategies. A similar study by Cuyckens et al. (1997) acknowledged the possibility that individuals may be sensitive to both broad *and* fine-grained distinctions, but in the absence of any data collected to examine this possibility, this remained a suggestion rather than a conclusion. To my knowledge, this possibility has not been addressed in polysemy literature. The significance of this gap in knowledge is highlighted by the fact that participants who sorted examples of these four words could not always be classified simply as lumpers or splitters. Instead, the data suggest that participants typically demonstrate both tendencies.

3.7.3 The use of "mixed-sense" groups

While at least one mixed-sense group – that is, a group comprised of sentences that I would judge to use both spatial senses and non-spatial senses of the target words –

was created in each of the four tasks, the tendency for participants to use such groups varied. Moreover, there was an unexpected pattern in the use of mixed-sense groups. On the one hand, most participants who sorted examples of *over* and *above* created mixed-sense groups. On the other, they were rare amongst the groups created for *under* and *below*. We might tentatively conclude from this outcome that the spatial vs. non-spatial distinction has varying degrees of semantic fundamentality across words and across participants.

This outcome contrasts with similar empirical studies undertaken by Sandra and Rice (1995) and Cuyckens, Sandra, and Rice (1997). Those studies, which saw participants sorting examples of a number of prepositions, resulted in data suggesting that the major distinction between examples of these words was whether or not they describe spatial relations, an outcome which the authors conclude to indicate separate mental representations for spatial and non-spatial meanings of these prepositions. Perhaps due to the space limitations imposed on these publications, the authors do not describe any individual participants' responses that suggest that this boundary was not always observed. However, such an outcome seems like an important one, worthy of at least brief mention, which suggests that a mixed-sense sorting strategy was not used.

When we talk about the words *over, under, above* and *below*, we tend to describe them as spatial words, and we contrast their spatial senses with non-spatial senses. The results of this study suggest that while the spatial/non-spatial distinction is often a very meaningful one, some participants judge this distinction not to be sufficiently meaningful to license two groups, and that there is a more important aspect of their meaning that overshadows this distinction. Some participants seem to have, therefore, senses of these words which are sufficiently flexible to describe spatial and non-spatial domains. Existing work on the polysemy of spatial words – both theoretical (e.g., Tyler and Evans, 2001) and empirical (e.g., Cuyckens, Sandra, and Rice, 1997) – highlights the distinction of spatial and non-spatial senses, a distinction no doubt influenced by theories of metaphor which predict that concrete experience (in this case, spatial configurations and our real or imagined interactions with them) structure our understanding of abstract phenomena (in this case, non-spatial meaning such as quantities and hierarchies) (e.g., Kövecses, 2002). That is

not to say that the relationship between spatial and non-spatial senses of these words is overlooked; on the contrary, it is explicitly argued that these non-spatial senses are understood on the basis of our experience with a particular spatial relationship. This position is illustrated particularly well in Tyler and Evans' (2001, p. 746) semantic network of *over*.

While research in polysemy acknowledges the relationship between concrete and abstract meanings, that is as far as it goes: there is a relationship, but the two domains remain distinct and are represented by distinct senses. In this task, we would expect this to manifest in the form of distinct spatial versus non-spatial groups. The data collected and presented are therefore difficult to reconcile with existing empirical research in the field, and existing theoretical accounts of the polysemy of spatial words, particularly those influenced by theories of metaphorical relations between concrete and abstract phenomena. These theories neither predict nor seem able to account for the transcendence of the spatial/non-spatial divide evident in participants' senses. These data might be better accounted for in terms of an exemplar-based model of categorisation. This model assumes that the features or characteristics of stimuli exist in a multidimensional space, and that these characteristics are represented by axes. It predicts that when a particular characteristic acts as the variable by which stimuli are categorised, e.g., colour, the corresponding axis expands to reveal enhanced diversity in the shapes of the stimuli, resulting in the similarity of colour of the stimuli decreasing. At the same time, axes corresponding to irrelevant dimensions, e.g., shape, shrink, which results in the relative similarity of the shape of the stimuli increasing.

While these data are clearly difficult to account for under existing accounts of polysemy, which borrow from prototype categorisation to explain the divisions and relations between uses, there is as yet insufficient evidence that an exemplar-based account of polysemy can better explain word sense representation. However, an early indication that this model may explain sense representation, taking into consideration the comments made above concerning the effect of the relevance of stimulus feature/characteristic axes, comes from participant O4's creation of an 'instead of' group, featuring the stimuli shown in Table 16.

Table 16 Members of participant O4's 'instead of' group

Stimulus sentence
Yeah can you turn that OVER please.
I turn it OVER
Turn that steak OVER, it's burning!
They should have voted Labour OVER Lib Dem
They live OVER the other side of the city
Get those cars OVER there
I think I'll sit OVER here
I want that one OVER there
bring the turning allowance of pelmet fabric OVER, and press with an iron to bond together

In this case, the members of the group are judged to be distinct from other stimuli by virtue of the fact that they all share a characteristic of opposition. I offer here a crude suggestion of how an exemplar-based model of polysemy and word senses might explain the creation of this group. Given the shared notion of "opposition", the axis/axes corresponding to the characteristic(s) of opposition expand to reveal the extent to which those characteristics are present in all of the stimuli. Irrelevant characteristics, such as the degree to which the stimuli described a spatial configuration, shrink, resulting in stimuli describing spatial and non-spatial scenarios becoming more similar to the point where they are indistinct from each other on that dimension.

In summary, there appear to be individual differences in whether participants find the distinction between spatial and non-spatial uses of these words to be sufficiently meaningful to divide them into two groups, and tendencies for creating "mixed-sense" groups varied across the four words studied.

3.7.4 The temporal or spatial nature of TEXT USES of *above* and *below*

We observed that participants who sorted examples of *above* typically sorted TEXT USE examples with sentences that were non-spatial in nature, or used labels that described temporal relations. In contrast, participants in the *below* task sorted TEXT USES with spatial items, or labelled their groups in spatial terms.

This unexpected dichotomy might be explained through analysis of what states *above* and *below* in text contexts refer to. The use of *above* to refer to a statement made in a text certainly indicates its physical position. However, it also directs the

reader's attention to a temporal position and, more specifically, to a location in their own temporal experience. The opposite cannot be said of *below*. While *above* and *below* in this context both underspecify the precise spatial and temporal location of the referent text, that *below* locates the forthcoming text in the future means that the exact temporal position of the referent text is less likely to be known than when *above* is used. While the spatial and temporal positions *above* pinpoints are rather vague, the pinpoints have already been experienced. Our understanding of *above* in these contexts to capture a temporal meaning is therefore sound: assuming we have read the text from the beginning, in a linear fashion, we will have encountered the referent text, and so will have experienced it temporally as well as spatially. In contrast, when we encounter *below* in these contexts, we know that there is potential for us to encounter it at some point in time, but that we may not necessarily encounter it – we may give up reading before we encounter the referent text, for example. However, its spatial location remains regardless of our continued interaction with the text. This may justify our understanding of this sense of *below* as one which is fundamentally spatial in nature.

In summary, there was an overarching tendency for participants to categorise TEXT USES of *above* in a manner which suggested they considered them to have a non-spatial, and possibly temporal, meaning. In contrast, participants in the *below* task tended to categorise these examples in a manner which indicated that they judged this sense to describe a spatial meaning.

3.8 Conclusions

This study begins to add to our understanding of the nature of the senses of polysemous words, specifically, the degree to which native speakers of English agree with each other about which uses of *over, under, above* and *below* share a meaning, and which uses do not. Analysis of agreement values, calculated using Morey and Agresti's adjusted Rand, suggest that participants fail to reliably agree with each other about how uses of these four words should be categorised when the categorisation criterion is the meaning of the word in context. Further evidence gathered from a more qualitative analysis supported this quantitative approach. Specifically, I observed that there were considerable differences in the number of groups individual participants created, indicating differences in the level of

granularity participants apply when considering whether examples of a given word are the same or different. I also observed differences in the use of "mixed-sense" groups, whereby some participants created groups containing what I judged to exemplify spatial *and* non-spatial uses of the target word. In contrast, analysis of how TEXT USES of *above* and *below* were classified revealed striking consensus about their meaning; while these uses of *above* were categorised in a manner suggesting that they have a temporal meaning, classification of these uses of *below* suggest that speakers judge them to have a spatial meaning.

We have observed that the senses of the polysemous words *over, under, above* and *below* do not appear to be shared by all speakers. Inter-participant disagreement over what constitutes an example of a particular sense has been observed in computational linguistics research (e.g., Passonneau, Baker, Fellbaum, and Ide, 2012; Passonneau et al., 2010). In this study the extent to which pairs of participants agreed with each other varied tremendously from pair to pair, and across the four words. While some disagreement might be reasonably expected, the extent of disagreement we have seen, and the fact that this disagreement is itself variant, pushes that expectation to the limit. However, the scale of the task may be to blame for poor inter-participant agreement. After all, if the significant cognitive demands the task imposes caused inconsistent or incoherent sorting decisions *within* participants, we should not expect their sorting decisions to be consistent *across* participants.

With this caveat, we can tentatively conclude that there may be individual differences in word senses. This outcome, paired with the findings reported in Chapter 2 concerning naïve participants' agreement with my sense distinctions, encourages linguists intending to pursue an analysis of a polysemous word that relies in whole or in part on their intuitions to proceed with caution.

This is an exciting outcome with theoretical and technological implications, and for this reason this conclusion demands further study. It was noted above that disagreement amongst participants may be a result of fatigue, boredom, or forgetting, as a result of the large scale of the task. To eliminate this possibility, a reduced and more structured iteration of the task, will be used to test further for individual differences in word senses. This study will be reported in the next chapter.

My urging of a rigorous approach to testing this outcome comes from my recognition that the scale of the inter-participant differences reported here is not currently accounted for in theoretical treatments of polysemy, nor in technological approaches. Researchers in word sense disambiguation studying polysemy tend to focus on developing a gold standard inventory of word senses. Creating a gold standard is intended to allow machines to correctly disambiguate the language produced by any native speaker of the target language. In the face of data that suggests that people disagree with each other about what the senses of a given word are, it seems that an interesting line of research might be to establish whether tailored word sense inventories are desirable and feasible. Gold standard and tailored sense inventories have mutually exclusive trade-offs: a gold standard system would allow any example of text or spoken language to be disambiguated to an acceptable standard, but disambiguation may not be perfect. A hypothetical tailored sense inventory would be intended to reach 100% accuracy when disambiguating a target person's spoken or written language, but, if individuals do differ in their sense distinctions, this algorithm would not achieve perfect accuracy with non-target individuals' language.

Chapter 4 Experiment 3: Testing an exemplar model of word senses

The sentence sorting task reported in Chapter 3 found disagreement amongst participants over how examples of *over, under, above* and *below* should be categorised when the sole sorting criterion is their meaning. As noted in the concluding remarks, the large degree of disagreement might be explained in methodological terms; specifically, the large scale of the task might have resulted in one or a combination of fatigue, boredom, confusion or semantic satiation. This possibility is easily addressed by reducing the scale of the task. The study reported here achieves this. In addition, a minor refinement of the methodology allows other, equally interesting issues to be addressed. Specifically, it was noted that the individual differences that were observed in Experiment 2 could be accommodated if we assume an exemplar-theoretic representation system for word senses. Accordingly, this chapter offers a refinement of Experiment 2; the scale is smaller and the stimuli more structured. In addition, the study collects data which is intended to shed light on whether word senses are stored in memory. Finally, the study aims to assess whether a central prediction of the (Generalised) Context Model of Classification (Medin and Schaffer, 1978; Nosofsky, 1986), namely *selective attention*, is observed in categorisation decisions made in these sorting tasks. Following convention in cognitive linguistics, the GCM will henceforth be referred to as the exemplar model.

Applying a model of categorisation developed in cognitive psychology to linguistic categorisation is not new in cognitive linguistics; after all, given the cognitive commitment, it makes perfect sense to assume that categorisation principles in other cognitive domains apply to linguistic categorisation. To date, linguistic categorisation accounts of polysemy have canonically assumed that a representation system inspired by Rosch and colleagues' prototype model can account for their storage and organisation (Brugman and Lakoff, 2006 [1988]; Tyler and Evans,

2001). More recent accounts have moved away from that model, and considered whether the exemplar model, a theory compatible with usage-based accounts of language, can account for the representation of word senses (Gries 2015).

This chapter is organised into two parts. The first part situates the study in a broader context and offers a critical review of research on word sense storage, contrasting exemplar- and prototype-theoretic accounts of word sense representation. Following my review of exemplar-theoretic accounts of word sense storage, I note that while exemplar theory has been drawn upon to account for word sense representation, to date this work has not tested the central prediction of exemplar theory – selective attention, which was introduced in section 1.9.2.1. While selective attention in linguistic categorisation has received relatively little focus, both selective attention and attention more generally have been the topic of study by linguists, and I dedicate some space in this literature review to introducing some work that has studied the relationship between (selective) attention and language.

Part 1: Literature review

4.1 Are word senses stored?

The question of whether or not word senses are stored has been given three answers. The first is that they are not. According to Ruhl (1989), many ambiguous words have a single, underspecified meaning. Context is used to substantiate the meaning, resulting in ad-hoc senses that are created on the fly, as and when necessary. Word meaning is, in his view, a product of linguistic and extralinguistic context (p. vii), and is pragmatic rather than semantic (p. ix). The second answer, which Murphy (2007, p. 57) describes as the "logical opposite" of a monosemy account is that every sense is stored in memory. This account may be compatible with an exemplar-based model of word sense storage: by virtue of storing every instance of a particular word in an organised, multidimensional space, word senses on the most minute scale – down to single exemplar senses – are stored. The third answer is that some are, and some are not. Tyler and Evans (2001, p. 727) are proponents of this position, claiming that there is a distinction between senses that are stored in memory, versus those that are mere "interpretations" created on an ad hoc basis.

4.1.1 Exemplar-based accounts of word senses

In line with this third, intermediate proposal, Murphy claims that "[p]olysemous senses will sometimes be explicitly represented and sometimes not, depending in part on how distant a given sense is from other, established senses" (2007, p. 65). This middle ground, "limited listings" proposal assumes that our encounters with a given word are mapped to a position in a multidimensional space. If a word is used in a similar way to how it has been encountered previously, it will be mapped to an already-known location in the space. If it is used in a novel way, a new mapping, to a previously-unencountered location, will be created. In this account, Murphy (2007) argues that two factors can cause multiple, distinct representations: frequency, and semantic distance – or similarity. However, he notes that these factors are related: if two uses of a given word are quite similar in meaning, they may initially be represented singly. However, if these similar uses are more frequent, the speaker is given more opportunities to establish what makes each use different, a process which can, over time, result in the separation of the two uses into distinct senses. In contrast, the more dissimilar two uses of a given word are, the fewer encounters one needs for the uses to be represented separately.

However, if two uses – let us call them sub-senses here – of a given word are very similar and very frequent, then one might predict that their high frequency could also result in their confusability: as well as being given more opportunities to establish what makes the two sub-senses contrast, the speaker is given more opportunities to establish what makes them alike. Further, it seems likely that examples of sub-senses *a* and *b*, i.e., a^1 and b^1 will be differentially similar to one another than examples a^2 and b^2 which in turn will be differentially similar to one another than examples a^3 and b^3. In other words, examples of sub-senses *a* and *b* will vary in how much they differ; some examples of *a* and *b* will be very dissimilar, while some will be very similar. In terms of a semantic space, some will be very proximal, while others will be very distant. In the absence of a clear divide between examples of sub-senses *a* and *b*, we might expect that they would have a single representation, as illustrated in Figure 38.

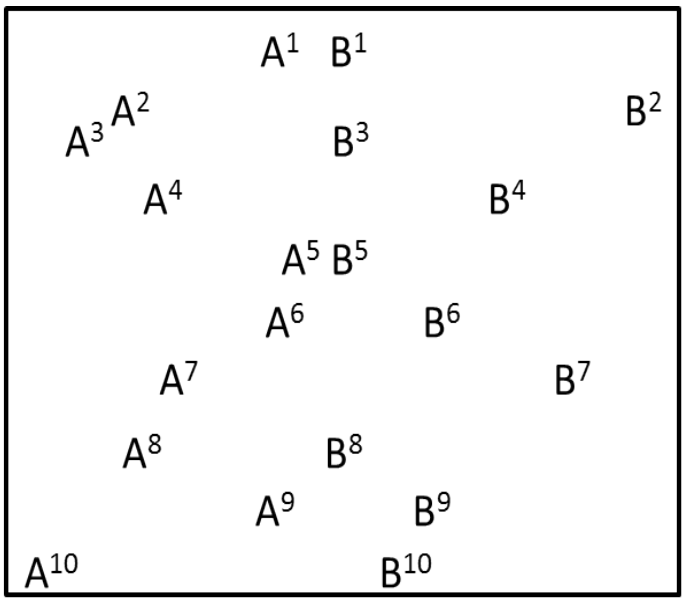


Figure 38 Diagram illustrating a single representation resulting from differentially-similar exemplars of sub-senses *a* and *b*

If these sub-senses are very frequent, then in order to create the two separate representations that Murphy predicts it is necessary for the more dissimilar pairs of examples to be much more frequent than the similar pairs. This is illustrated in Figure 39.

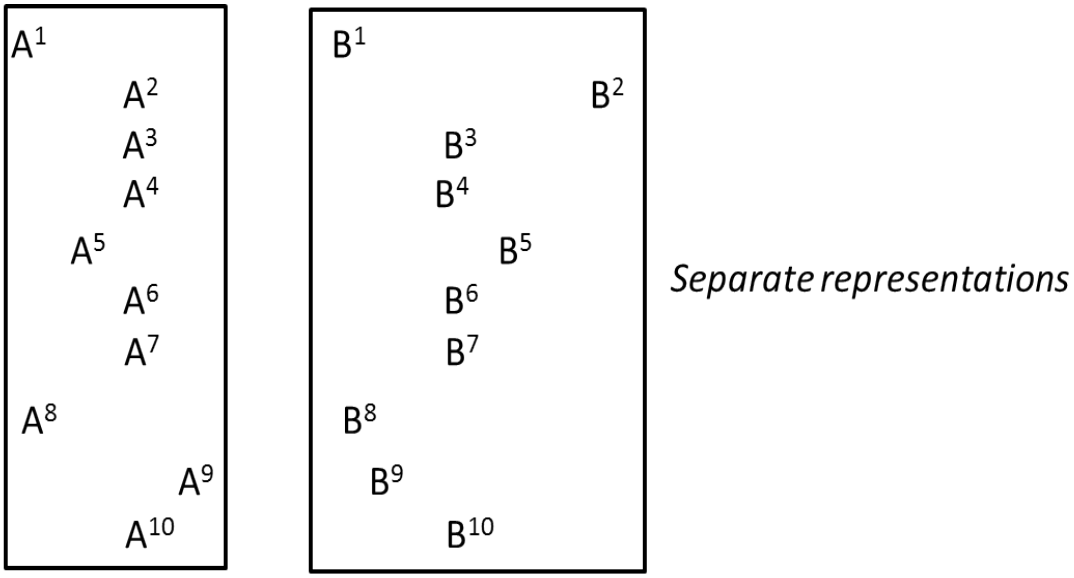


Figure 39 Diagram illustrating two separate representations resulting from multiple dissimilar exemplars of sub-senses *a* and *b*

Since we are talking here about the effect that similar pairs of examples of two sub-senses have on the explicit representation of (sub-)senses, the notion that lots of dissimilar examples of pairs of sub-senses are necessary to support this aspect of the

proposal seems a little illogical. Rather than creating distinct senses, and assuming that word senses *are* represented in this "limited listings" manner, I would instead propose that frequent examples of similar sub-senses would result in a much broader, coarse-grained sense (encompassing *a* and *b*) which can, should the need arise, be inspected more closely to identify the finer distinctions separating *a* and *b*. This is illustrated in Figure 40.

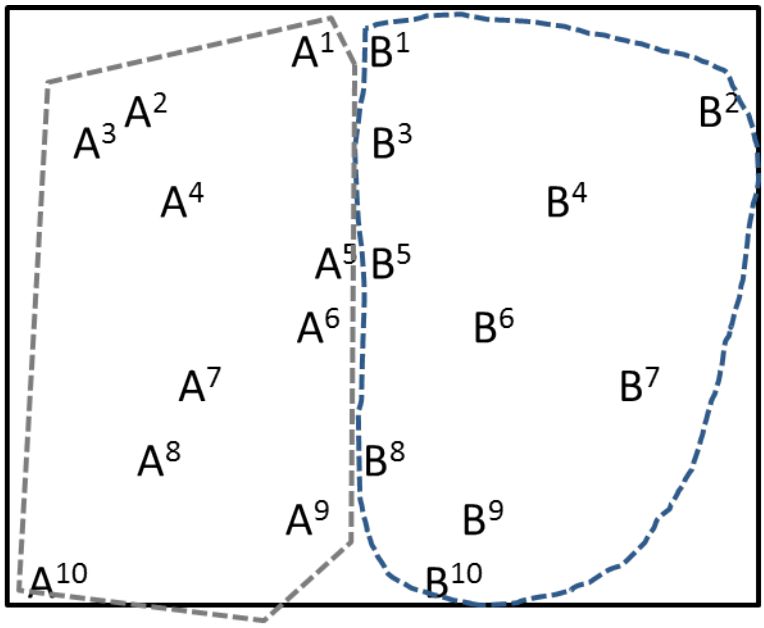


Figure 40 Diagram showing original single representation captured in Figure 38 (black outline), plus finer-grained separate representations that can be accessed as needed (grey and blue dashed outlines)

This proposal is consistent with Ide and Wilks' (2007, p. 66) comments about the relative necessity of coarse- to fine-grained senses. They propose that "'sense disambiguation' is really a process of step-wise sense refinement that progressively distinguishes 'sub-senses' *as needed for understanding*" (my emphasis). In other words, when finely-grained senses are not necessary, they are not accessed. A multidimensional space model, in which sub-senses are represented, does allow fine distinctions to be accessed where necessary.

There is a clear overlap between the language Murphy (2007) uses and the language used by proponents of exemplar models of categorisation. In both cases, exemplars – linguistic or otherwise – are located in a multidimensional space, in which the dimensions are properties held by the objects. The exemplar model, which has been

shown to account well for categorisation of non-linguistic phenomena, appears to be compatible with an intuitive middle-ground account of sense enumeration along the lines Murphy proposes. Because the representation consists of exemplars, Murphy's proposal also appears to be compatible with the usage-based account of language that cognitive linguists support. Under this model, differences in exposure to a given word may explain why humans may fail to agree with each other about what fine-grained sense of a target word is used in a stimulus sentence. Applying this to computational linguistics, given that automated WSD algorithms are trained using vast corpora of language that may not necessarily match the exposure of the person who created the "gold standard" sense inventory, a model such as this may also explain why automated WSD does not result in perfect agreement.

While Murphy's use of particular terms *suggests* that his account may be compatible with an exemplar-theoretic model of word sense representation, other scholars explicitly align themselves with the exemplar model when describing the psychological status of word senses. For example, Gries (2010), in his discussion of the behavioral profile approach introduced by (Divjak 2003) explicitly claims that the model is compatible with exemplar theory (p. 340). Of most relevance to the present research is that the model assumes that exemplars of a given word are tagged for particular information, and that this information corresponds to a dimension in a multidimensional psychological space. In this way, Gries and his colleagues' approach undoubtedly aligns with the multidimensional space assumed in exemplar theory, and it certainly accommodates why exemplars of a category are more similar to each other than they are to exemplars of other categories.

4.1.1.1 Attention and language

While exemplar-theoretic models of word sense representation have been proposed, none, to my knowledge, discuss the effect of selective attention, a concept of central importance in exemplar theory that was introduced in section 1.9.2.1. Before reviewing the literature on selective attention in linguistic categorisation, I will first offer an overview of attention more generally, and note that while *selective* attention is generally overlooked in cognitive linguistics, attention has long been a topic of interest in the field. Its longstanding place in cognitive linguistic research may be explained on the grounds that attention is a domain-general cognitive function.

A simple example of seeing and hearing an ambulance while driving demonstrates the domain-general nature of attention: one can visually attend to the direction in which the ambulance is headed, and aurally attend to the sound of the siren, and the driver's interpretation of the resultant visual and aural signals can allow them to determine whether it is headed toward or away from them, and alter their driving path accordingly. This example also indicates that attention is not equally distributed, but can be moved and refocused according to internal and external demands and influences. For example, when listening to a piece of music, one listener might listen closely to the lyrics, overlooking other aspects such as timbre or the bass line. In contrast, another, perhaps someone trying to learn how to recreate this piece of music themselves, might attend closely to the chords, ignoring all other aspects for now. The psychological processes captured by the umbrella term of attention are incorporated into cognitive accounts of a range of linguistic phenomena, from our initial acquisition of words, to how those words are organised in a sentence. At the start of our lives, attention – specifically, joint attention – is understood to play a pivotal role in our acquisition and development of language. Joint attention, in which two or more people "(1) attend to the same object, (2), know that the other does, and (3) know that the other knows" (Verhagen, 2015, p. 283), is understood to be a crucial factor in language acquisition style and lexical development (e.g., Tomasello and Farrar, 1986; Tomasello and Todd, 1983). This ability to share, and indeed direct, attention provides an early foundation for successful communication, which Langacker (2001, p. 144) proposes to arise when a speaker and hearer direct and focus their attention in coordination, resulting in attention being directed to and focused on the same "conceived entity." Once these first words are acquired and from that point onwards, linguists and psychologists propose that attention "appear(s) to constrain directly the use of linguistic structures during language production" (Myachykov, Tomlin and Posner, 2005, p. 351).

The way we use language therefore reflects our general tendency to distribute attention unequally (Myachykov, Tomlin and Posner, 2005). For example, the way we talk about events reflects an equal balance of attention between agent and patient. As outlined by Myachykov et al. (2011. p. 99), a rich body of literature has demonstrated the impact that attention on either agent or patient in a transitive event has on a participant's description of that scene; briefly, when participants initially

see an agent character, they almost always use the active voice to describe the event. When they see the patient character, use of the active voice declines, and the passive is chosen by some participants[6]. Likewise, the way we describe entities reflects the level of attention one pays to their features. As the two examples below indicate, the level of specificity in a description will be determined by the degree of attention the author (and/or reader) is likely to pay to the object.

8. This is a 19th Century Victorian mahogany bookcase with a galleried rectangular top above open front of 3 adjustable shelves flanked by stiles mounted with corbels and raised on plinth base. (LoveAntiques 2017)
9. White oak bookcase for sale in great condition small damage to top corner but an easy fix cost over 300 (Gumtree 2017)

4.1.1.2 Selective attention in language and linguistic categorisation

As the previous section demonstrated, attention and its relationship with language have been and continue to be an area of interest to cognitive linguists. This fact, alongside the fact that categorisation – including exemplar theories thereof – is also a central topic in cognitive linguistics, makes it all the more surprising that *selective* attention has received sparse attention by linguists, including those who explore exemplar-theoretic accounts of linguistic categorisation. In the following section, I will introduce some research that has examined the role that selective attention plays in language. The majority of this research considers selective attention out of categorisation contexts, and instead addresses the role that selective attention plays in normal and second language development. Following this, I will then move to address selective attention's status in cognitive linguistics as an aspect of exemplar theory that is generally overlooked in exemplar-theoretic accounts of linguistic categorisation. Finally, I will introduce an example of research by Kalyan (2012) that *has* considered selective attention and its role in linguistic categorisation. This particular piece of research is especially important as it offers an answer to a question typically left unanswered: if categorisation depends on selective attention, exactly what aspect(s) of the object to be categorised should one attend to? His

[6] As research on languages other than English has demonstrated, however, attention is not the sole determinant of word order choice (Myachykov, Garrod, and Scheepers, 2011).

answer is offered as part of a wider study of differential acceptability of questions with long distance dependencies.

4.1.1.2.1 Selective attention and language

As noted in the preceding section, selective attention and its role in linguistic categorisation have come under relatively little scrutiny, despite the growing currency of exemplar theory to account for linguistic categories. That said, a large body of work does exist to examine the role that selective attention plays in language more generally, and especially in the development of first and additional languages. Scholars and clinicians working with populations with specific language impairment (SLI) have observed that children with SLI demonstrate different attention abilities, indicating the importance of selective attention for language acquisition. Stevens, Sanders, and Neville (2006) observed in an ERP study that while typically developing children responded more positively to probes in attended stimuli than in unattended stimuli, children with SLI responded in a similar way to probes in attended and unattended stimuli. The fact that the two groups responded similarly to probes in unattended stimuli, but differently to probes in attended stimuli indicates that children with SLI show deficits in signal enhancement rather than distractor suppression. In other words, selective attention deficits in children with SLI do not appear to be caused by an inability to overlook distracting, irrelevant stimuli, but instead by an inability to enhance target, relevant stimuli. Likewise, selective attention has been proposed to play an important role in second language acquisition, due to the suggestion that L2 acquisition depends on the ability to consciously attend to the form of input (Robinson 1995). Indeed, Lively et al. (1993) has proposed that one cause of the difficulty of acquiring a second language is the fact that "retuning" one's selective attention mechanisms, to allow one to attend to features of a stimulus that are irrelevant in the native language but which are relevant in the second language, is a difficult task.

Elsewhere, research in adult language impairment has opened up debate on the direction of the relationship between selective attention and language. Lupyan and Mirman (2013) studied aphasic and matched control participants' performance in low-dimensional and high-dimensional categorisation tasks. In low-dimensional tasks, stimuli shared one or few common features and were required to be organised

into narrow categories such as “things that are orange”. High-dimensional tasks featured stimuli that shared multiple features and were to be organised into broad categories such as “farm animals”. The authors proposed that low-dimensional categorisation requires greater cognitive control in that participants must attend to one or only a few features of an array of stimuli, and ignore others that might distinguish them. In this way, the selective attention demands created by low-dimensional tasks were much greater than those presented by high-dimensional tasks. They further hypothesised that cognitive control resulting in successful low-dimensional categorisation may be supported by language – in particular, naming abilities. The authors observed that aphasic participants performed significantly worse than matched controls in low-dimensional categorisation tasks, but not on high-dimensional categorisation tasks. This was found to be irrespective of the lesion site, indicating that rather than categorisation *and* naming impairments being jointly impaired by cognitive control mechanisms, naming ability supports low-dimensional categorisation. In other words, the authors propose that language may support selective attention.

4.1.1.2.2 Selective attention and linguistic categorisation

According to exemplar-theoretic accounts of categorisation, selective attention is what produces categories, formed of exemplars represented in a multidimensional space, and what facilitates accurate categorisation. By selectively attending to a particular feature of a stimulus, such as its colour, exemplars are reorganised in the space according to their colour. Red objects, for example, are located closely together, and separately from black objects. If, on the other hand, the same objects were being organised by their shape, black and red triangles, which in the first scenario were in different categories, will now find themselves in the same category, and separate from black and red circles. In this way, categories are inherently dynamic. This notion of dynamic categories logically entails that categories are not fixed representations. The possibility that categories are not fixed, but are created ad hoc, entails that any exemplar-theoretic model of word senses must acknowledge that word senses are also impermanent structures, and simply *potential* categories of exemplars. To my knowledge, this possibility has not been addressed or explored in cognitive linguistic work on word senses. The notion of word senses as potentials has, however, been alluded to in lexicographic literature. Kilgarriff, in his rejection

of word senses entirely, proposed that "word senses exist only relative to a task." (1997, p. 1). While Kilgarriff does not position this statement in an exemplar-theoretic context, his claim is compatible with the theory.

Given the centrality of this effect to the exemplar model of categorisation, and the growing currency of exemplar-theoretic accounts of word sense representation, this is an important gap. It is somewhat surprising, though, that this gap exists to begin with. After all, observing selective attention effects in word senses would be a strong indicator that they are represented in a manner consistent with exemplar theory. There is evidence, however, that in some cases linguistic categorisation is treated separately from categorisation of other phenomena. For example, in her exemplar-based account of the representation of constructions, Bybee (2006, p. 716), says that "In some versions of exemplar representation, exemplars are scattered randomly through space. Only when categorization of a new exemplar is necessary are they organized by similarity". Bybee states here that in a categorisation task (experimental or real), the position of exemplars in space changes from random to organised; this is consistent with the account central to Medin and Schaffer and Nosofsky's models. However, the notion that they are "randomly scattered" is at odds with the notion of a multidimensional space. Exemplars are located in a location corresponding to values on multiple dimensions, i.e., in a location that is anything *but* random. Afterwards, Bybee states that "Because linguistic categorisation takes place so often I propose that linguistic categories […] are more entrenched in the sense that frequently used categorisations have an impact on neurological organisation." (p. 716). I take this to suggest that Bybee judges exemplar categories of linguistic phenomena, such as phonemes and constructions, to have a fixed location in an organised space, which is therefore not modulated by selective attention. Furthermore, categories, under this account, are understood as fixed representations. In this way, although advocating an exemplar-based account of linguistic categorisation, Bybee appears to treat linguistic categorisation separately from categorisation in general. This approach is not unique; in his overview[7] of

[7] This overview evaluates three models of inflectional morphology based on exemplar models of categorisation: Albright and Hayes' (2003) Minimal Generalization Model, Daelemans et al.'s (2002) Memory-Based Learning model, and Skousen's (1989, 1992) Analogical Model. He observed that even incidental variables (i.e., features corresponding to dimensions in a multidimensional

computational models of inflectional morphology, Chandler (2010) notes that models that have incorporated "weak" versions of the generalized context model developed by Nosofsky (1986), in which feature weightings and therefore selective attention are removed, perform worse than models which do include some form of feature weighting. Elsewhere, a limited amount of research has studied – and found – selective attention effects in linguistic categorisation (Ellis, 2006; Francis and Nusbaum, 2002; Lively et al., 1993).

It is unclear whether omission of selective attention and/or feature weighting from linguistic accounts of categorisation is intentional, or comes as a result of a misunderstanding of exemplar theory. However, within his overview, Chandler (2010) claims that accounts of exemplar categorisation in cognitive linguistic literature are fundamentally inaccurate, in that they claim that categories (rather than only exemplars of a *potential* category) are represented in the model; this is certainly true of Bybee's proposal. He attributes this, in part, to Nosofsky's (and presumably therefore also to Medin and Schaffer's, 1978) invocation of a multidimensional space, stating that the proximity of exemplars in a psychological space like this is understood therefore to reflect relative similarity, and therefore category membership. "Clouds" of exemplars, to use Pierrehumbert's (2000) term, are therefore understood in cognitive linguistic literature to represent a category. Chandler takes exception to this, arguing that these clouds do not necessarily represent categories. I make no comment on this argument, but do challenge his claim that it is Nosofsky's proposal of a multidimensional space that has resulted in inaccurate understanding of the nature of the exemplar model by cognitive linguists. Both Medin and Schaffer's (1978) context theory of classification and Nosofsky's (1986) generalized context model make absolutely plain the role of feature weightings and selective attention in categorisation decisions. In Medin and Schaffer's (1978, p. 212) words,

> Selective attention can be represented by changes in the salience or similarity parameter for dimensions. That is, the similarity parameter of

---

psychological space) associated and stored with exemplars have been demonstrated to significantly improve inflection choices. In light of this, and due to their absence in the Minimal Generalization Model, he ultimately concludes this model can be rejected (p. 406).

> two cues along a dimension is less when that dimension is attended than when it is not.
>
> This assumption is designed to capture the consequences of active hypothesis testing. For example, if subjects were trying out the possibility that all red stimuli belong to Category A and all green stimuli to Category B, they might code much less information about other attributes such as size or form than otherwise. As a result, the effective similarity of two size or two form cues might be greater than usual, and the effective similarity of red and green would be expected to be less than otherwise.

And in Nosofsky's (1986, p. 41) words,

> It is assumed that [...] multidimensional perceptual representation underlies performance in both the identification and categorisation paradigms. However, a selective attention process is assumed to operate on this perceptual representation that can lead to systematic changes in the structure of the psychological space and associated changes in interstimulus similarity relations. […] Selective attention is modelled by differential weighting of the component dimensions in the psychological space[…]. In geometric terms, the weights act to stretch or shrink the psychological space along its coordinate axes.

Selective attention in itself rules out the possibility that categories have fixed representations in these models. Given the clarity with which Medin and Schaffer (1978) and Nosofsky (1986) describe selective attention and feature weighting, I disagree with Chandler, and suggested instead that should a scholar opt to omit selective and attention and/or feature weighting, and therefore treat linguistic categorisation as a separate phenomenon, this is not the consequence of Chandler's perceived lack of clarity about the long-term position of exemplars in a multidimensional space.

If a scholar working within the cognitive linguistic framework *does* opt to omit selective attention from their account, this would be rather surprising. To do so would be to treat categorisation of linguistic versus other phenomena separately. This would be contrary to the cognitive commitment underpinning the cognitive linguistic framework: the commitment to explaining linguistic phenomena – in this case, linguistic categories – in terms of what we already know about cognition more generally – in this case, what we know about categorisation as a domain-general process. In the case of Bybee's statement, that the frequency of linguistic categories

necessitates that they be treated separately and understood to have more fixed representations, we should note that categorisation is by necessity a continuous process. As a result, we perpetually encounter stimuli which we are then tasked with categorising. Treating linguistic categorisation separately on the grounds that linguistic categories are frequent is therefore somewhat problematic.

Whether due to decisive omission or misunderstanding, to my knowledge, work that studies the prediction of selective attention effects in polysemy is absent. This leaves open an interesting opportunity to test this prediction empirically, and to assess whether a modified version of the exemplar model is necessary to accommodate linguistic categorisation.

4.1.1.2.3 What role can linguistic applications of exemplar theory play in theory development?

While exemplar-theoretic accounts of linguistic categorisation do not typically account for what role selective attention plays in categorisation decisions, there are exceptions, some of which were introduced briefly in the previous section. In this section I will introduce research by Kalyan (2012), who aims to account for differences in acceptability of questions with long-distance dependencies (LDD questions) in exemplar-theoretic terms; more specifically, he accounts for these differences in terms of selective attention. This piece of work is unusual not only for its foregrounding of the role of selective attention in exemplar-theoretic linguistic categories, but for the fact that it offers a suggestion to a question often left unanswered in exemplar theory research: how exactly does one determine what they should attend to in a categorisation event? He notes that when exemplar theory is invoked to account for linguistic categorisation, the feature(s) to which one should attend to decide whether a novel item is a member of a particular category are generally left unspecified; moreover, he notes the tendency for scholars to assume that similarity is invariant. In line with the notion in exemplar theory that categorisation is context-specific, Kalyan proposes that the feature(s) to which one should attend must be determined in the context of the specific categorisation event; in other words, when presented with the same stimulus on two different occasions, the features that one must attend to will vary according to what the categorisation goal is. Kalyan proceeds to propose that the features that one should attend to are typically those that tend to characterise the category features themselves; i.e., one

should attend to "necessary features" of the category. Kalyan's answer to the question of what features one should attend to was provided following his aim to reconcile Dąbrowska's (2008) and Ambridge and Goldberg's (2008) ostensibly competing accounts of why some verb seems less acceptable than others in questions with long distance dependencies (LDD questions). He observes that Dąbrowska's proposition that LDD questions featuring verbs of *saying* and *thinking* are considered more prototypical depending on how similar the main verb is to *think* or *say*. Ambridge and Goldberg, he notes, do not reach for a similarity-based explanation, and instead propose that the restriction of certain verbs arises from the potential incompatibility between the information structure properties of the verb and the construction (p. 543). Specifically, they propose that that the constituent forming the gap in a filler-gap construction must not be backgrounded, and instead must either be the topic of the non-gapped clause or within its potential focus domain. As a result, an LDD question is judged as more acceptable if its matrix verb is a "light" or "bridge" verb such as *think* or *say*, than if it is a manner-of-speaking verb such as *mumble*, or a factive verb such as *notice*, due to the fact that the complement clauses of manner-of-speaking and factive verbs are typically backgrounded, while the complement clauses of light or bridge verbs need not be.

While Dąbrowska and Ambridge and Goldberg offer ostensibly competing accounts of why some verbs, but not others, are acceptable in LDD questions, Kalyan suggests that the two can be reconciled. He proposes that Ambridge and Goldberg's proposition that acceptability results from information structural compatibility between construction and verb can be explained in similarity terms, if one considers that the acceptability of an LDD question is a function of how similar a matrix verb is to the matrix verbs that appear in attested LDD questions in terms of how far the verb foregrounds its complement clause. However, he notes that this account might be objected to on the grounds that in order to successfully judge the acceptability of an LDD question – or indeed produce an acceptable LDD question – one needs to know that its acceptability hinges on this very specific property of one component of the construction. He proposes that this knowledge may emerge through exposure to and usage of the construction, noting that the high frequency of *think* and *say* as the main verb in LDD questions, and the fact that both verbs strongly foreground their complement, allows speakers to gradually come to know that the acceptability of an

LDD question is likely to depend upon the degree to which the verb in the construction foregrounds its complement.

As well as demonstrating the explanatory power of selective attention in categorisation decisions, Kalyan's paper serves the demonstrate the important role that linguistic applications of theories of cognition more generally can play in developing those theories. Given the vast role that categorisation plays in language, the study of linguistic categories offers a rich testing ground for evaluating the credibility of categorisation theories, for testing their generalisability, and for developing them.

4.1.2 Prototype-based models of sense storage

Murphy's model described in section 4.1.1 overlaps to some extent with exemplar-based models of concept categorisation. Other models of sense representation draw on alternative accounts of categorisation. Perhaps the most studied alternatives are those which draw on the prototype categorisation model proposed by Rosch and her colleagues during the 1970s (e.g., Rosch and Mervis, 1975). Brugman and Lakoff propose radial categories of polysemous words and their senses, with "a central member and a network of links to other members", and with "each noncentral member of the category [being] either a variant of the central member of ... a variant on a variant" (2006 [1988], p. 109). They argue in favour of a full-specification account of word sense storage, in which very finely-grained senses are stored. While they are finely-grained, they remain abstractions of particular instances of the target word. In their theory of Principled Polysemy, Tyler and Evans (2001) propose that a protoscene underpins a semantic network of polysemous word senses. The protoscene and senses are abstractions of particular instances of a given word, which are further specified through interpretation of the context in which they fall. They propose that some of these abstracted senses are stored in memory.

These models overlap with the arguments Rosch makes about the fuzziness of categories, varying degrees of membership, and the fact that abstractions rather than exemplars are stored. However, the radial category and Principled Polysemy model differ from the prototype model in that the central sense in both cases is not an

abstraction of the entire category, but is instead a very basic schema which underpins all other senses, and from which those senses can extend.

4.2 Sense storage: Conclusion

There is no firm answer to the question of whether or not word senses have fixed representations stored in memory. On the one hand, some scholars argue that word meaning is highly abstract in nature, and that this abstract meaning is substantiated by sentential and environmental context to allow successful disambiguation and communication. On the other hand, other scholars have claimed that at least *some* senses are stored in memory. In the absence of any firm conclusions either way, further investigation is therefore needed.

In addition, though, if we claim that word senses *are* stored, we must be able to explain, or at least hypothesise, *how* they are stored. Moreover, linguists operating under the cognitive commitment central to cognitive linguistics must offer an explanation that is compatible with what we know about other, non-linguistic categorisation. The linguistic categorisation of interest here – the categorisation of exemplars of polysemous words into senses – has been argued to be compatible with competing models of cognitive categorisation: exemplar-based models, and prototype-based models. While prototype-based models are commonplace in cognitive linguistic accounts of polysemy and word senses, studies that conclude that an exemplar-theoretic explanation is necessary are fewer in number. This fact is in contrast with the vast body of research that has been carried out on exemplar theory, and is at odds with the arguments suggesting that exemplar-theoretic models better account for categorisation observation than prototype-based models (Medin and Schaffer 1978; Murphy 2004, p. 103). It seems, therefore, that further research is necessary to understand whether word senses can be understood in exemplar-theoretic terms. One interesting means of testing this is to investigate whether the central prediction of the exemplar model – selective attention – is observed in a linguistic categorisation experiment. This approach would serve to inform our understanding of how applicable the exemplar theory is to linguistic categorisation. Moreover, given that the prototype model does not predict selective attention, any observations of selective attention effects would challenge a prototype-theoretic account of word senses. Beyond that, this approach also offers the opportunity to test

Bybee's proposal that linguistic categories are fixed, entrenched representations, and are not modulated by selective attention.

Part 2: Investigation

4.3 Aims

4.3.1 Looking beyond individual differences

While it was the sole purpose of the study reported in Chapter 3 to examine individual differences in word senses, sentence-sorting tasks are rather versatile. The research reported in this chapter therefore takes this opportunity to return to the matter of individual differences in word senses, before expanding beyond this topic to consider other aspects of the psychological status of word senses; namely whether or not there is evidence that they are stored in memory, and, if they are, whether their representation is compatible with the (generalised) context theory of classification developed by Medin and Schaffer (1978), and Nosofsky (1986), commonly referred to as the exemplar model.

4.3.2 Are word senses stored in memory?

Asking participants to complete the same sentence-sorting task twice, separated by a delay of two months, will address the matter of whether word senses are stored in memory. If participants agree how the sentences should be sorted with themselves better than with other participants, this will be taken as evidence that word senses are indeed stored. If there is no significant difference in how well participants agree with themselves versus how well they agree with other participants, or if there is variation in how well participants agree with themselves in comparison with how well they agree with other participants, this does not necessarily rule out the possibility that senses are stored. The exemplar model of categorisation argues that categorisation is task-based; i.e., we construct a category by adjusting the relative positions of known exemplars in a multidimensional space on an ad hoc basis. Given that the purpose of and instructions for the task will remain fixed each time participants complete it, it is not expected that the criterion/criteria participants use to categorise the stimuli will change. However, external factors such as different time constraints and changing interpretation of the instructions may result in changes in criteria, and consequently reduced consistency.

4.3.3 How are word senses stored in memory?

The representation of word senses in memory is the topic of ongoing debate, as noted in section 4.1. Within research that argues in favour of sense storage there is disagreement over the manner in which they are stored. The cognitive linguistic literature has traditionally viewed word senses as being stored in memory in a prototype-like representation system (e.g., Brugman and Lakoff, 2006 [1988]; Tyler and Evans, 2001). However, more recent accounts such as those by Murphy (2007) and work on behavioural profiles (e.g., Gries and Divjak, 2009), introduce the possibility that word senses may be more akin to exemplar categories, consisting of every previously encountered token of a particular word. The study reported here investigates the suitability of modelling word senses in terms of exemplar categories by testing a prediction of the exemplar model of categorisation, namely the effect of selective attention.

If we can view word senses as having an exemplar-like representation, we can expect to observe selective attention effects in a semantic categorisation task such as the sentence-sorting task used here. In the present study, I ask whether the presence of a highly distinctive difference in the meanings of the target words represented in the stimuli coincides with differences in sorting behaviours compared with how stimuli that do not feature this difference are sorted. Specifically, I will ask whether, when participants sort a set of sentences representing *both* spatial and non-spatial senses, they sort them differently to how a set of sentences representing *either* spatial *or* non-spatial senses are sorted.

When participants sort sentences that represent a mixture of both spatial and non-spatial senses, the presence of both sense types may create a contrast that serves to shrink other possible categorisation dimensions. Specifically, the presence of non-spatial uses alongside spatial uses may cause participants to attend to this broad distinction first (i.e., they will decide whether the sentence is spatial or non-spatial), and possibly then to other characteristics, for example, whether the sentence describes a two- or three-dimensional configuration.

The goal of categorisation is accuracy (Nosofsky, 1986), as proven by the fact that failure to accurately categorise the exemplar pictured below may have fatal consequences.

Figure 41 Lion, which may be a member of a number of categories such as ANIMALS ONE MAY SEE ON SAFARI, and PREDATORS (Wikipedia, 2016)

If there is a very obvious attribute that divides the stimuli, attending only to this dimension allows for an easy yet accurate (if somewhat coarse) categorisation solution. Put simply, participants might adopt a categorisation strategy whereby the key distinction criterion is whether or not the target word has a spatial or non-spatial meaning; after that, they can pursue further categorisation distinctions if they judge that more refined groups exist amongst the stimuli. In contrast, participants who categorise examples representing a single type of sense, either spatial or non-spatial, will not have this advantage. They must seek out finer, less obvious distinctions within the stimuli.

By randomly assigning participants to one of two conditions, the *single sense-type condition*, featuring only spatial, or only non-spatial senses, or the *mixed sense-type condition*, featuring a combination of spatial and non-spatial senses, I will be able to see whether there are systematic differences in the way sentences are sorted when the contrast of a different sense type is removed, as would be consistent with an exemplar account of word sense storage. Differences in sorting behaviours will be measured in terms of the number of groups participants in each condition create. An exemplar-theoretic account of word sense representation would predict that stimuli presented in both experimental conditions are sorted into fewer groups in the mixed sense-type condition than in the single sense-type condition.

4.3.4 What can networks tell us about word meaning?

Collation of sorting data gathered from a task allows one not only to explore questions around individual differences and representation of word senses, but also how individual examples of a target word are semantically related to each other. Similarity matrices created automatically in OptimalSort provide the starting point for exploring such relationships, but they allow a rather surface-level analysis, and it is not possible to measure the significance of the matrix. In this study I investigate the utility of using network visualisations of sentence-sorting data to study word senses.

Network analysis is conventionally associated with the study of complex systems and relationships therein. Outside of linguistics, they have been used, for example, to study social relationships, ecosystems, biological processes, and food webs (Newman 2012). Networks, an example of which is shown in Figure 42, consist of nodes, connected to each other by edges. For example, in a food web, nodes represent species in an ecosystem, while edges might represent predator-prey relations. Network visualisation algorithms can also produce communities, which Kauffman et al. (2014, n.p.) describe as "groups of nodes with dense intra-community edges and sparse inter-community connections." Communities therefore provide further detail about the structure of the network, and open up the possibility of finding previously unknown or unexpected modules (Blondel et al. 2008). In this research, nodes represent the stimulus sentences and edges represent a participant's decision to categorise them into the same group. Good et al. (2010, p.1) note that identification of communities within a network might provide a "principled way to reduce or coarse-grain a system by dividing global heterogeneity into relatively homogenous substructures." Following this, communities are understood to capture highly similar exemplars, and might be understood as senses.

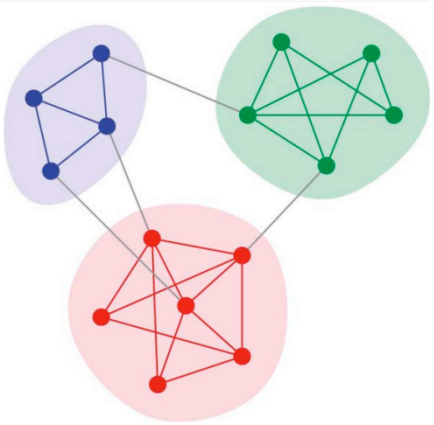


Figure 42 Nodes, edges and communities in a network (Newman 2012)

Associated with communities displayed in a network are modularity values, which quantify the reliability or significance of the communities identified by the

algorithm. In other words, a modularity value is a measure of the overall quality of the network. Values range from -1 to 1. A "good" set of communities has an attendant modularity value that is closer to 1, and will feature groups that have more internal connections than would be expected at random. A "bad" set of communities has a modularity value closer to zero, reflecting no more connections between community members than we could expect by chance (Good et al. 2010, p.1). However, they recommend caution in the interpretation of modularity values. Elsewhere, it has been informally suggested that modularity values are taken lightly (Levallois 2013). A means of testing the reliability of modularity values is to assess whether they correlate with other measures of reliability. In this research, agreement between participants over how stimuli should be categorised is measured statistically, using Morey and Agresti's adjusted Rand. This chapter will briefly ask whether or not there is a relationship between statistical agreement calculated for each task, and the modularity value of each associated network.

4.4 Research questions

In summary, the present chapter will explore the following questions in the manner described below:

Is there evidence that word senses are stored in memory?

This will be studied by assessing whether there is a significant difference between how similar an individual's sorting decisions are in the two tasks (*intra-participant agreement*) and how similar each individual participant's sorting decisions in the first task are with all other participants' sorting decisions in the second task (*inter-participant agreement*). I predict that if word senses do have some form of mental representation, participants will agree with themselves significantly better than with other participants, and that they will agree with themselves more often than with other participants.

Are sentence-sorting decisions subject to selective attention effects?

This will be measured by comparing the number of groups used to categorise stimuli that appear in both conditions. If an exemplar model can account for the representation of word senses, I predict that participants in the mixed sense-type

condition will sort the stimuli that appear in both conditions into significantly fewer groups than participants in the single sense-type condition do.

Is there evidence that participants have different senses of the target words?
This will be achieved by calculating the degree of agreement between participants, and by measuring the quality of communities identified by the network algorithm. These two sets of values will also be compared to assess whether they correlate with each other.

4.5 Data collection

In this section, I outline the approach I took to gathering data to answer the question of whether there is evidence of individual differences in word senses, and whether there is evidence of selective attention effects in the sentence-sorting tasks used. It then provides relevant information about the participants who completed the tasks. Thirdly, details of the materials used are provided, followed by detailed information about the task procedure. Finally, it specifies the statistical model and data visualisation tool used to analyse the data.

4.5.1 Methodology

The open-sort version of the sentence-sorting task, described fully in Chapter 3, was adopted here. All participants completed the task online using the OptimalSort system. Participants completed the same sorting task twice, separated by a period of two months. Participants were not made aware that the tasks were identical.

4.5.2 Participants

Following the study reported in Chapter 3, this research focuses on the decisions made by native speakers of English. In the present study, 205 native English speakers completed both parts of the study. Responses by participants whose native language was not English, or who did not complete both parts of the task, are not included in the following analysis. The sample represents a broad range of educational, geographical and occupational backgrounds. The majority (84%) of participants were born in a majority English-speaking country outside of the United Kingdom. Most (64%) had completed at least a bachelor's level qualification, and the highest concentration of occupations was in the professions (40%). However, all educational backgrounds, from incomplete secondary school-level education to doctoral qualification were represented, as were the majority of the occupations

recognised by the International Standard Classification of Occupations, which was used to classify participants' occupation responses.

4.5.3 Stimuli

The task stimuli consisted of 36 examples of one of *over, under, above* or *below.* Sentences were edited versions of extractions from the internet and from the spoken and written sections of the British National Corpus, edited to make the examples well-formed sentences. Where appropriate, the sentences were edited to reduce their length. By reducing their length, I could ensure that as many sentences as possible were visible on the screen at any one time.

Three sets of stimuli were used for each of the four target words. One set, sorted by participants in the mixed sense-type condition, consisted of the mixture of spatial and non-spatial uses used in the closed-sort task reported in Chapter 2. Participants in one of the two single sense-type tasks either sorted a set consisting of sentences representing what I judged to be exclusively spatial senses of the target word, or those representing only non-spatial senses. The stimuli used can be found in Appendix 5.

Stimuli in the single sense-type condition comprised the spatial or non-spatial uses used in the mixed sense-type condition, topped up with additional spatial or non-spatial uses to make a total of 36 sentences. These sentences were selected and edited as described above. In total, the 36 sentences represented what I judged to be six examples of six distinct senses. The use of the same spatial or non-spatial uses as used in the mixed sense-type condition in the single sense-type condition was a conscious decision: it allows direct comparison of how the same stimuli are categorised in different conditions, thus revealing whether selective attention determines categorisation decisions.

For all but the *over, under* and *above* tasks, the stimuli in the single sense-type condition tasks included six exemplars of three senses used as stimuli in the mixed sense-type task, i.e., exemplars of three non-spatial senses, or exemplars of three spatial senses. In the case of the *below* single sense-type task, the spatial stimuli included six exemplars of four senses used as stimuli in the mixed sense-type task.

The stimuli used in the *below* non-spatial task comprised of two groups of six sentences from the mixed sense-type task stimuli. In all cases, these stimuli were topped up with new sentences to result in a total of 36 stimuli. The reason for this difference between the origins of the stimuli for the *over, under,* and *above* tasks on the one hand, and the *below* task on the other hand, is as follows. As noted in Experiment 1, I judge what I term TEXT USES of *below,* such as *Fill in below all the tasks that you do in a typical day*, to have a non-spatial, temporal meaning. However, as noted in Chapter 3, participants in an open-sort task reached a different conclusion, and appear to judge them to have a spatial meaning. Given that the purpose of the single sense-type task was to see how sorting decisions changed when a highly salient semantic contrast – the spatial/non-spatial distinction – was absent, I decided to recognise the TEXT USES as examples of a spatial sense, and no longer as a non-spatial sense, and included these uses in the spatial stimuli.

4.5.4 Procedure

Prior to starting the task, information about the task and how responses would be stored was presented on the screen. Participants were required to read and confirm that they understood this, and provide their informed consent to participate in the study. Participants were then shown written instructions about how to complete the task (see appendix 6 for a copy of these instructions). Briefly, participants were instructed to sort a set of sentences into one or more groups. They were explicitly instructed to sort sentences based on what the capitalised target word meant in each sentence. It was made clear that the goal of the task was to sort all of the sentences into groups in which the meaning of the capitalised target word was the same in each member of the group. The instructions disappeared when the participant moved the first sentence, but could be recalled at any time.

After being presented with the task instructions, the task was revealed on the screen. Sentences were presented in a column on the left side of the screen, with a larger blank sorting pane to the centre and right of the screen. Participants were advised to read all of the sentences, considering carefully what the target word, which was shown in full capitals, meant in each sentence. Afterwards, they were required to move each sentence into the sorting pane to create a group, after which further sentences could be dragged and dropped into it. Sentences could be moved in and

out of categories until the participant was satisfied with their sorting decisions. Once they were satisfied with their groups, participants were asked to give each group a label describing the meaning of the target word captured by the constituent sentences. Once they had completed this, they clicked a button to indicate that they had finished the tasks. The results were then stored and accessible in the back end of the programme. Participants were required to sort all of the stimuli before they could submit their responses.

4.5.5 Statistical analysis

Data collected in this task were statistically analysed for agreement using Morey and Agresti's adjusted Rand (Morey and Agresti, 1984). As discussed in section 4.3.4, the data were also run through a network visualisation algorithm to create networks based on the sorting decisions, and to quantitatively measure the quality of the networks.

4.6 Results and discussion

4.6.1 Are word senses stored in memory?

4.6.1.1 Do participants categorise sentences in the same way in each task?

Table 17 shows the average extent of intra-participant agreement in each word and condition, along with the sample standard deviation revealing the extent of variation around the mean. The values therefore represent how well, on average, the participants in each word and condition agreed with themselves about how the stimuli should be sorted at time 1 and time 2 (hereafter T1 and T2). We can see some tendencies emerging: with the exception of the tasks for *over*, participants had highest intra-participant agreement when sorting stimuli that represented both spatial and non-spatial senses. In addition, with the exception of the tasks for *below*, participants changed their sorting decisions most when sorting spatial uses of these words.

Table 17 Descriptive statistics of intra-participant agreement values

	Mean intra-participant agreement	SD
Above non-spatial	0.749	0.170
Above mixed	0.773	0.205
Above spatial	0.733	0.189
Below non-spatial	0.581	0.198
Below mixed	0.752	0.212
Below spatial	0.685	0.151
Over non-spatial	0.840	0.105
Over mixed	0.832	0.093
Over spatial	0.645	0.263
Under non-spatial	0.561	0.180
Under mixed	0.804	0.122
Under spatial	0.504	0.256

On the whole, a range of average intra-participant values are returned; from 0.504 for *under* spatial, to 0.84 for *over* non-spatial. While we are not necessarily concerned with obtaining optimal agreement, but are interested in studying agreement levels for their own sake, it is interesting to note that only participants in the *over* non-spatial and *under* and *over* mixed sense-type tasks agreed with themselves to an extent that we can describe as acceptable, reaching an average intra-participant agreement score of at least 0.8 (Neuendorf, 2002, p. 3). The participants in the three *above* tasks are consistently just below this boundary, but amongst the other tasks agreement is classified, according to Neuendorf, as reflecting great disagreement.

When participants are asked to complete two identical sentence-sorting tasks, there is, therefore, evidence of variation across the tasks in how well participants agree with themselves over how the stimuli should be sorted when there is a delay of two months between them. As well as that, though, there is variation *within* the sample of participants in each pair of tasks. At one extreme, there is low variation around the mean intra-participant agreement score for *over* mixed and *over* non-spatial, but very high variation in the *over* spatial and *under* spatial tasks. We can see, therefore, that it is not simply the case that participants' sorting decisions remain differentially constant over time depending on the stimuli that they are sorting, but that there is sometimes very wide variation in how consistent participants' sorting decisions are, regardless of the stimuli they are asked to sort.

On the whole, then, there is variation in intra-participant agreement values between words, within words, and between participants. Of the twelve pairs of tasks, only three produce mean intra-participant agreement values in excess of the lower acceptable bound endorsed by Neuendorf (2002). The implications of this finding for the question of whether there is evidence that word senses are stored in memory are complex. On the one hand, the presence of some low mean intra-participant agreement scores indicates that participants do not reliably sort the same stimuli into the same groups when completing identical sorting tasks separated by a period of two months. Consistently high intra-participant agreement scores would give a firm indication that the categories participants produce, which I tentatively assume to represent word senses, are stored in memory. There is indeed evidence of very good agreement in some tasks. How is it that some pairs of tasks result in very high mean intra-participant agreement scores, and others very low scores? The presence of both high and low intra-participant agreement values, both within and across tasks, is difficult to explain under models which posit either zero word sense storage (Ruhl 1989), or full storage of word senses (Tyler and Evans, 2001). An intermediate model in which word senses have *some form* of mental representation may explain these different findings.

4.6.1.2 Do participants agree with themselves more than they do with other people?

The previous section presented data that indicates that word senses are stored in some form in memory. This interpretation can be further tested. If word senses *are* stored, we would expect participants to have better agreement with themselves about how the stimuli should be sorted than with other participants, and that they should reach a higher level of consensus with few participants. This hypothesis is tested in two ways: first, a paired-samples t-test is used to study whether there is a significant difference between intra-participant agreement values and the mean inter-participant agreement value for each participant. Second, I examine how regularly participants' intra-participant agreement scores rank higher than their inter-participant agreement scores. Given the uneven sample size in each task, inter- and intra-participant agreement scores were transformed into percentile ranks.

The results of the t-test indicate that, on the whole, participants agree with themselves significantly better than they do with other participants: $t(204) = 12.516$, $p = .000$, $d = .75$. Further, intra-participant agreement values typically ranked in at least the 75th percentile (M = 75.18, SD = 28.855). Taken together, these values strongly indicate that participants' sorting decisions are more similar to their own than to those of other participants.

If we consider this finding in conjunction with the findings that there is variation – and sometimes low levels – of intra-participant agreement, it is difficult to make firm conclusions concerning whether or not word senses are stored in memory. However, in and of themselves poor intra-participant agreement values do not rule out the possibility that word senses are stored in memory. In fact, the fact that high agreement values can be found in the means, and amongst individual participants, in addition to the finding that participants on the whole agree with themselves significantly better and more often than they do with other participants, indicates that they may indeed be stored in memory, though perhaps not in a fixed, permanent form. The exemplar model is able to account for both the low and high agreement scores found here. In the model, categorisation is task-based; as noted in section 4.3.3, the multitude of dimensions on which a particular exemplar sits vary in their length according to the task at hand. The characteristics of the stimuli an individual participant selectively attends to in the first task might vary from those she attends to in the second task. Accordingly, the characteristics of the stimuli she attends to will vary, thus resulting in the length of the associated dimensions changing. When dimensions change, the proximity of exemplars will also change. As a consequence, the categories she identifies will be different.

I propose that so long as an individual participant selectively attends to the same characteristic in each task, and therefore adopts the same sorting strategy each time, they will construct categories that are very similar, simply because the task has remained the same. If, second time around, their sorting strategy changes, the task changes, resulting in changes to the categories they construct. The characteristics they selectively attend to, and therefore their sorting strategy may change for any number of reasons; they may wish to complete the task more quickly, they may

notice and attend to details in the stimuli that they had overlooked in the first task, or they may interpret the instructions differently.

In summary, the presence of some high intra-participant agreement values indicates that word senses may be stored in memory, as long as we understand them as categories generated in response to task demands. This interpretation is consistent with Kilgarriff's claim that "word senses only exist relative to a task." (1997, p. 1). I interpret these data, therefore, in terms of the exemplar model, and argue that these findings provide early support for application of the model to account for the representation of polysemous word senses. The possibility that the exemplar model can account for word sense representation will be explored further in the following two sections.

4.6.1.3 Interim conclusions

The fact that individuals vary in how consistent their sorting decisions are when they complete identical sentence-sorting tasks two months apart is difficult to explain, should we wish to propose that word senses are stored in some fixed and stable form in memory. Some participants display a very high degree of consistency in their sorting decisions, which suggests fixed representations. However, some participants made significant changes to their categorisations each time they complete the task, which suggests no representations. The fact that all but one participant's sorting decisions were more consistent than would be expected by chance is promising, but not particularly meaningful. A cognitive model of word senses that posited that word senses are not stored in memory in any form would surely predict that, in a pair of sorting tasks such as these reported here, participants' ability to recognise the similarity between pairs or groups of uses of a target word would not disappear altogether, thus forcing participants to categorise the stimuli on by their best guess. However, the intra-participant agreement values are often quite a way from chance level (i.e., an agreement value of 0), and a generous proportion (42%) have agreement scores of at least 0.8, which is reckoned as a sensible cut-off point for acceptable agreement (Neuendorf, 2002). Likewise, if word senses were not stored in any form, we should not expect individual participants to agree any better with themselves than they do with other participants. The opposite outcome was found, with participants agreeing with themselves significantly better than with other

participants, and with intra-participant values typically ranking in at least the 75th percentile.

It appears instead, then, that word senses may have some form of representation in memory. As noted previously, the range of agreement values is accommodated by an exemplar-theoretic account of categorisation. However, further investigation is needed to assess whether, if word senses *are* stored in memory in any form, the exemplar model can best account for their representation. The following sections will address some of the explicit and implicit predictions of the model: that categorisation is inherently task-based, and therefore similarity of exemplars is decided relative to a task; and that categorisation of sentences is subject to individual differences.

4.6.2 Are sentence-sorting decisions subject to selective attention effects?

If principles of the exemplar model are applicable to the data collected in this study, there should be measurable differences in sorting decisions according to what broad semantic categories are represented in the stimuli. In this research, participants *either* categorised sentences which only used either spatial or non-spatial uses of the target word (i.e., the single sense-type condition), *or* they sorted stimuli which represented a combination of spatial and non-spatial uses (i.e., the mixed sense-type condition). In the mixed sense-type condition, there is a major semantic contrast that divides the stimuli into two distinct categories. I predict that when participants read the set of sentences prior to beginning to sort them, this distinction will be salient. Accordingly, I predict that the spatial/non-spatial distinction will at least be an initial categorisation criterion. In an exemplar model, because that major point of contrast exists, it should be the case that finer distinctions that might divide spatial uses and non-spatial uses into smaller groups are overlooked. In the absence of this major point of contrast, i.e., in the single sense-type condition, fine distinctions will not be overshadowed. As a result, where stimuli appear in both the single sense-type and mixed sense-type condition tasks, I predict that there will be more distinctions amongst these stimuli in the single sense-type condition.

4.6.2.1 Number of groups used to categorise sentences that appear in single and mixed sense-type conditions

Table 18 below shows the mean number of groups created to categorise stimuli that appear in both task conditions.

Table 18 Mean number of groups used to categorise sentences that are found in both single sense-type and mixed sense-type task conditions

Task	Mean number of groups T1	Mean number of groups T2
Above spatial in single sense-type task	3.50	3.25
Above spatial in mixed sense-type task	2.12	2.12
Above non-spatial in single sense-type task	3.08	3.46
Above non-spatial in mixed sense-type task	2.96	3.15
Below spatial in single sense-type task	4.00	4.19
Below spatial in mixed sense-type task	2.93	3.07
Below non-spatial in single sense-type task	2.75	2.83
Below non-spatial in mixed sense-type task	1.47	1.73
Over spatial in single sense-type task	4.19	4.25
Over spatial in mixed sense-type task	3.75	3.81
Over non-spatial in single sense-type task	3.83	3.92
Over non-spatial in mixed sense-type task	3.19	3.50
Under spatial in single sense-type task	3.47	3.47
Under spatial in mixed sense-type task	2.25	2.40
Under non-spatial in single sense-type task	3.50	3.70
Under non-spatial in mixed sense-type task	3.15	3.20

An independent samples t-test measured whether the number of groups participants created to categorise stimuli that appeared in both task conditions was different. I predicted that participants would categorise the stimuli into more groups in the single sense-type task. This pattern was found. At T1, in the single sense-type tasks an average of 3.61 groups were used (SD = 1.14), versus an average of 2.71 groups (SD = 1.15) in the mixed sense-type tasks. At T2, the sentences were sorted into an average of 3.7 groups (SD = 1.08) in the single sense-type condition, versus an average of 2.84 groups (SD = 1.12) in the mixed sense-type condition. This difference was significant at both T1 ($t(280)=6.59$, $p = .000$, d = .79) and T2 ($t(280)=6.46$, $p=.000$, d= .77).

These findings therefore indicate that the presence of a significant semantic contrast results in participants sorting stimuli into broader groups. Under an exemplar model, attention to the semantic contrast dividing the stimuli serves to shrink other potential dimensions corresponding to other characteristics that might be used to produce finer distinctions. To exemplify this, let us consider the distinctions in the same set of stimuli made by participant BMi9 in the mixed sense-type condition, and BS10 in the spatial condition, as shown in Table 19 below.

Table 19 Spatial sentences and categorisation decisions made by participants BMi9 and BS10

Sentence	Participant	Category label used by Participant BMi9	Participant	Category label used by Participant BS10
BELOW the crags a well-built tunnel could be seen.	BMi9	physically under	BS10	physically under
BELOW the front windows the extension was divided into two sections.	BMi9	physically under	BS10	vertically underneath on a flat surface
Don't paint BELOW the windowsill.	BMi9	physically under	BS10	vertically underneath on a flat surface
Fill in BELOW all the tasks that you do in a typical day.	BMi9	lower on the page	BS10	further along in the text
Give us your fun verdict by dialling the numbers BELOW.	BMi9	lower on the page	BS10	further along in the text
I called out to the people on the beach BELOW, but they didn't hear me.	BMi9	physically under	BS10	physically under
I pinned my name badge BELOW the logo on my tshirt.	BMi9	physically under	BS10	vertically underneath on a flat surface
In the situations listed BELOW identify what your information needs would be.	BMi9	lower on the page	BS10	further along in the text
Instead of being up high the box was down BELOW.	BMi9	physically under	BS10	further along in the text
It will be argued BELOW that economic reconstruction was a success.	BMi9	lower on the page	BS10	further along in the text
Look at the sentence BELOW, what does it say?	BMi9	lower on the page	BS10	further along in the text
Serve with Sharp sauce (see BELOW).	BMi9	lower on the page	BS10	further along in the text
She had a mole just BELOW her right eye.	BMi9	physically under	BS10	vertically underneath on a flat surface
Tabith stood BELOW, watching him.	BMi9	physically under	BS10	geographically lower
The campus was shrinking BELOW me into a collection of children's play houses.	BMi9	physically under	BS10	geographically lower
The crocodile sank BELOW the surface.	BMi9	physically under	BS10	layer underneath
The people in the flat BELOW wouldn't stop shouting.	BMi9	physically under	BS10	physically under
The sleeves gradually get tighter and end BELOW the elbow.	BMi9	physically under	BS10	layer underneath
The walk provides wonderful views of Mallerstang BELOW.	BMi9	physically under	BS10	physically under
There are two iron rings on the wall BELOW the painting.	BMi9	physically under	BS10	vertically underneath on a flat surface
They established an iron foundry in the valley BELOW the church in 1790.	BMi9	physically under	BS10	South
We dredged BELOW the mud at the bottom of the river.	BMi9	physically under	BS10	layer underneath
When we got BELOW the next layer the concentrations became stronger.	BMi9	physically under	BS10	layer underneath
Your mates are down BELOW, watching you.	BMi9	physically under	BS10	geographically lower

This table shows that in the case of these two participants, stimuli that appear in both conditions are categorised very differently. Participant BMi9 uses just two categories to classify all 24 sentences, making a distinction between sentences that they label 'lower on the page' (referred to elsewhere in this thesis as TEXT USES), and all other sentences, which are captured by a fairly catch-all 'physically under' label. In contrast, participant BS10 observes six distinctions within the group. They identify the same TEXT USE sentences (though include an additional sentence which BMi9 classified elsewhere), but amongst the other sentences identify much finer distinctions.

In this case, it appears that the presence of a spatial/non-spatial contrast made differences *within* spatial sentences and non-spatial sentences less salient. Assuming that the sentences and their categorisations can be modelled as occupying a multidimensional space, when this significant contrast is present, distinctions along other dimensions corresponding to characteristics that might otherwise be used to distinguish between exemplars shrinks, making exemplars which might otherwise be judge to differ along that dimension closer together, thus making them appear more similar along that dimension. In contrast, when that significant semantic difference is absent, as was the case in the single sense-type tasks, finer distinctions between the meanings of the target words can be detected.

4.6.2.2 Interim conclusions

This study tested the hypothesis that when a significant semantic contrast is present amongst stimuli, participants will categorise stimuli into fewer groups than when that semantic contrast is absent. The data and subsequent statistical analysis strongly support this hypothesis. This finding corresponds to a central prediction of the exemplar model, which predicts that categorisation is task-based and categorisation decisions will vary according to the sorting parameter on which the categorisation strategy is based. The findings reported here can stand alone as evidence in support of the conclusion that word senses are stored in memory as potential categories of previously encountered exemplars. This conclusion is further supported by evidence noted earlier suggesting that word senses may be represented in some form in memory. These findings jointly allow the conclusion that word senses may be understood as *potential* categories of exemplars, in the way that exemplar

categorisation proponents argue that non-linguistic concepts are represented (Medin and Schaffer, 1978; Nosofsky, 1986). I therefore move away from the proposal made in other cognitive linguistic accounts of polysemy that polysemous words themselves are categories (e.g., Tyler and Evans, 2001; Brugman and Lakoff, 2006 [1988], Taylor, 2003). I also conclude that the exemplar model appears to be able to account for their representation, rather than a prototype-based model advocated in previous research (e.g., Tyler and Evans, 2001; Brugman and Lakoff, 2006 [1988]).

4.6.3 Individual differences in word senses

4.6.3.1 Do participants agree with each other about how the sentences should be sorted?

If we base our judgment of the degree of inter-participant (dis)agreement in terms of Neuendorf's (2002) argument that agreement values of less than 0.8 indicate great disagreement, the mean agreement values shown in Table 20 indicate that, on the whole, participants fail to reach an acceptable level of consensus about how the stimuli should be sorted.

Table 20 Descriptive statistics of inter-participant agreement values at time 1 and time 2

Time 1					
	Mean	**SD**	**Min.**	**Max.**	**Range**
Above non-spatial T1	0.614	0.159	0.202	1	0.798
Above mixed T1	0.59	0.212	0.01	1	0.99
Above spatial T1	0.619	0.132	0.349	0.919	0.571
Below mixed T1	0.493	0.212	0.061	1	0.939
Below spatial T1	0.517	0.131	0.216	0.814	0.598
Below non-spatial T1	0.457	0.172	0.143	0.927	0.784
Over non-spatial T1	0.735	0.156	0.379	1	0.621
Over mixed T1	0.787	0.11	0.472	0.994	0.523
Over spatial T1	0.495	0.196	0.002	0.872	0.87
Under non-spatial T1	0.455	0.162	0.02	0.849	0.829
Under mixed T1	0.629	0.151	0.315	1	0.685
Under spatial T1	0.249	0.161	-0.022	0.698	0.72
Time 2					
	Mean	**SD**	**Min.**	**Max.**	**Range**
Above non-spatial T2	0.63	0.132	0.329	1	0.671
Above mixed T2	0.653	0.166	0.195	1	0.805
Above spatial T2	0.665	0.134	0.361	0.947	0.585
Below mixed T2	0.446	0.194	-0.012	0.937	0.949
Below spatial T2	0.524	0.133	0.253	0.881	0.628
Below non-spatial T2	0.675	0.147	0.362	1	0.638
Over non-spatial T2	0.767	0.13	0.447	1	0.553
Over mixed T2	0.742	0.115	0.418	1	0.582
Over spatial T2	0.45	0.217	-0.015	0.822	0.837
Under non-spatial T2	0.466	0.147	0.161	0.896	0.735
Under mixed T2	0.65	0.15	0.283	1	0.717
Under spatial T2	0.246	0.156	-0.013	0.684	0.697

Further, the standard deviations reveal that pairwise agreement values are, relative to the mean, really rather variant; indeed, there is also variation in *how* varied agreement is across the tasks. The 'Min.' and 'Max.' columns further demonstrate this variation. Let us look at an example, referring to the cell highlighted in red. In the case of the mixed sense-type task for *below* at T1, the mean agreement value is 0.446, but there is a case in which a pair of participants completing this task achieve a much lower level of agreement – indeed, agreement lower than we would expect

by chance: -0.012. At the other end of the scale, in the same task one pair of participants achieved very close agreement in their sorting decisions, with an agreement score of 0.937, as shown in the cell highlighted in green. Indeed, at least one instance of perfect or near-perfect agreement is observed in the majority of the 24 tasks. On the other hand, there are cases in which some pairs of participants reach agreement *less* than would be expected by chance. We therefore observe two striking outcomes. First, if we understand the categories participants create to classify the stimuli to reflect the senses they find meaningful, there is evidence of individual differences in word senses. Second, the degree to which pairs of participants agree with each other is variant, ranging from exceptionally high to exceptionally poor agreement.

Differences in how the stimuli should be sorted in this task correspond to the differences observed in sorting decisions seen in Experiments 1 and 2. The possibility that different speakers have different senses of polysemous words was introduced in my analysis of the first, closed-sort study. At that point, poor agreement with my sense distinctions was taken to suggest that the senses that I find meaningful do not correspond to those held by other speakers. It was acknowledged, though, that it might simply have been the case that the participants in that study agreed upon the senses of those words, and simply disagreed with me. The results of this study and the open-sort tasks reported in Chapter 3 undermine this conclusion, and support the alternative interpretation that not only do participants disagree with my word senses, but that they also disagree with each other about which exemplars of these polysemous words share the same meaning. It was noted, however, that poor inter-participant agreement in Experiment 2 might have been the result of methodological issues. That task was on a very large scale, with participants sorting 100 sentences. It was speculated that the scale of the task might have caused one or a combination of fatigue, boredom, confusion or semantic satiation. These factors may have resulted in sorting decisions that were not always consistent and coherent. Without sorting strategies that are executed systematically, we can expect nothing *but* weak agreement. This task addressed this possibility by reducing the scale of the task; unlike in Experiment 2, participants only sorted 36 stimuli. The stimuli were also edited to ensure that the sentence was well-formed and relatively compact. There was also additional structure imposed on the stimuli, in that I classified them

into six exemplars of what I judged to be six distinct senses. Such structure was not imposed in Experiment 2. Despite improvements to the structure of the stimuli and task design, poor average agreement values remain. The larger number of participants in this study compared with the open-sort study in Experiment 2 (205, versus 44), makes the interpretation that there are individual differences in word senses more robust.

The findings complement existing work that has found individual differences in linguistic phenomena. For example, Dąbrowska and Street (2006) found evidence that correct interpretation of the passive construction is related to participants' level of education. A range of research has found evidence of individual differences in milestones in and routes to language acquisition (Bates, Dale, and Thal, 1995). In a recent series of studies of interpretations of an ambiguous temporal metaphor, Duffy and colleagues have gathered evidence that individuals with different personality types, and different types of job, interpret metaphors differently (Duffy et al, 2014; Duffy and Feist, 2014; Duffy, 2015). The data gathered in this study provide further support for the notion that individuals differ not only in their linguistic ability, but also in their linguistic representations.

4.6.3.2 What can networks tell us about individual differences in word senses?

Network visualisations offer the potential for qualitative and quantitative analysis of sorting decisions across groups of participants. Qualitative analysis permits the study of communities that emerge in the network, and relationships (or absence thereof) between communities. Networks can also be studied quantitatively; of particular interest to this study is the potential for quantifying the "quality" of the network, based on the network's modularity value. A network structure that perfectly represents every participant's categorisation decisions will have a modularity value of 1. A network structure that corresponds less well to all participants' decisions will have a lower modularity value. In this study, we might predict that networks with low modularity values will be generated using data which has low inter-participant agreement. The utility of modularity values has, however, been informally challenged (Levallois, 2013). Comparison of network quality against agreement values offers the chance to see whether or not modularity values are indeed useful.

Table 21 shows the mean inter-participant agreement value for each task, along with the modularity value of the associated network. Networks generated for each task are provided in Appendix 7.[8] These values show that the quality of networks produced was highly variant. Pearson's *r* test was used to investigate whether there was a relationship between the extent of inter-participant agreement in each task and the quality of the associated network, as measured by its modularity score. This produced a strong and significant correlation ($r = .953$, $p = .000$).

Table 21 Mean level of inter-participant agreement and modularity value of the network produced for each task

	Mean agreement value	Modularity score
Above non-spatial T1	0.614	0.514
Above non-spatial T2	0.630	0.537
Above mixed T1	0.590	0.427
Above mixed T2	0.653	0.483
Above spatial T1	0.619	0.425
Above spatial T2	0.665	0.456
Below non-spatial T1	0.457	0.279
Below non-spatial T2	0.675	0.484
Below mixed T1	0.493	0.385
Below mixed T2	0.446	0.355
Below spatial T1	0.517	0.423
Below spatial T2	0.524	0.452
Over non-spatial T1	0.735	0.636
Over non-spatial T2	0.767	0.674
Over mixed T1	0.787	0.659
Over mixed T2	0.742	0.642
Over spatial T1	0.495	0.375
Over spatial T2	0.450	0.317
Under non-spatial T1	0.455	0.325
Under non-spatial T2	0.466	0.371
Under mixed T1	0.629	0.46
Under mixed T2	0.650	0.432
Under spatial T1	0.249	0.142
Under spatial T2	0.246	0.159

When based on the sentence-sorting data I have collected, network modularity values measure how accurately a network represents the groups all participants created; i.e., it tells us how well the network's communities correspond to all of the

[8] Due to the size of these files, they are best viewed on a computer screen.

participants' groups. The fact that modularity values and mean agreement scores correlate so strongly is therefore not particularly surprising; if participants make different sorting decisions, which is quantified by the agreement scores, a network will not be able to accurately represent every participants' groups, but will instead produce something akin to an "average" of the groups all participants created. So while this outcome is not particularly remarkable, it does indicate that modularity values, in this study at least, are meaningful measures. Indeed, if we were to argue that acceptable agreement is represented by an agreement score of at least 0.8, as Neuendorf (2002) does, we might also argue that an acceptable modularity value is approximately 0.67. This relationship is an original contribution to knowledge about the meaningfulness of modularity values.

Further research is necessary to investigate the utility of network visualisations of sentence-sorting data. It seems counterintuitive to attempt to glean information about the similarity of exemplars of polysemous words based on their positions in a low quality network; after all, these low quality networks fail to adequately represent the sorting decisions made by all participants. However, networks of both high *and* low quality may allow us to understand more about how, in the face of individual differences in word senses, communication nonetheless proceeds successfully. Let us return to Ide and Wilks' (2007) comment that coarse-grained senses may be the most useful level of meaning for successful disambiguation. It might be that smaller communities in high quality networks can be collapsed to form the coarse-grained senses that Ide and Wilks advocate. Equally, the larger, poorly-differentiated communities in low quality networks may be understood as coarse-grained senses. This idea is discussed in more detail in the following chapter.

4.6.3.3 The theoretical implications of individual differences in word senses

As discussed in sections 4.6.1.3 and 4.6.2.2, there is some evidence that word senses may indeed be stored in some form in memory, if we understand their storage in terms of an exemplar-theoretic categorisation system. The finding that individuals seem to disagree about which exemplars of a given word share the same meaning suggests that the semantic categories held by these participants, taken as word senses, is consistent with the exemplar model. The exemplar model predicts that categorisation is task-based, meaning that categories are structured in response to a

categorisation strategy determined by the characteristic the individual selectively attends to, such as categorisation by colour, shape, smell, and so on. While all participants received identical instructions, the crux of which specified that they should categorise the sentence according to the meaning of the capitalised target word, the strategies individual participants used to categorise the sentences is likely to be subject to some variation. Some participants were satisfied with broad, coarse categories, whereas others split the stimuli into smaller groups reflecting finer distinctions. Equally, participants may selectively attend to different aspects of the meaning of the target word. For instance, in the second *above* non-spatial task, participant AM18 classified together sentences that are judged by others to reflect positions on qualitative *and* quantitative scales. For this participant, the use of *above* to mark a position on a metaphorical scale was enough to categorise such exemplars as a member of the same category. In contrast, participant AM13 created a category for exemplars that used *above* to describe a position on a strictly numerical scale. Accordingly, if they judged a sentence to capture this meaning, it was assigned to this category. If it described the position on a metaphorical but *not* quantitative scale, it was assigned to a different category.

In the previous sections I have argued that word senses are represented in memory in a way that is compatible with the exemplar model of categorisation. Uncovering individual differences in the sense categories participants created is unproblematic in this model. The model claims that categorisation is task based, meaning that if the task (as manifest by a particular sorting strategy or set thereof) is different across individuals, then the categories they create will likewise differ.

4.6.3.4 Interim conclusions

Analysis of inter-participant agreement scores has revealed three key findings. First, the data have shown that pairs of participants' sorting decisions are rarely identical, and that on the whole, there is widespread disagreement. Second, while average agreement values paint a fairly poor picture of how much consensus participants as groups reached, pairwise agreement values reveal instances of both perfect agreement and agreement lower than would be expected by chance. This indicates that some participants agree with each other more than others. These findings are compatible with all of the findings discussed concerning word sense storage, and the

exemplar model, which I have recommended as a model of word sense representation. Finally, by comparing the modularity value, or numerical "quality" value of networks produced using sentence-sorting data with mean agreement values for each associated task, the utility of the modularity score has been supported, and a suggestion for what might constitute an "acceptable" network solution has been offered.

4.7 General discussion

In this chapter I set out to answer three questions:

1. Is there evidence that word senses are stored in memory?
2. Are sentence-sorting decisions subject to selective attention effects?
3. Is there evidence that participants have different senses of the target words?

A set of twelve open sentence-sorting tasks, each completed by participants on two occasions separated by a delay of two months, was used to answer these questions. The magnitude of inter- and intra-participant agreement was calculated statistically using Morey and Agresti's adjusted Rand. The study aimed to investigate not only whether word senses are stored, but whether their storage is compatible with the exemplar model that is gaining increasing currency in cognitive linguistics. The second question addresses a central prediction of the model, which states that categorisation is task-based, and that categories of stimuli will vary according to the categorisation criterion/criteria the categoriser selectively attends to. Under this account, categories are claimed to consist of previously-encountered exemplars. On this basis, in conjunction with the argument that categorisation is task-based, categorisation decisions are expected to display individual differences.

Intra-participant agreement values were used to assess the degree of consensus participants reached with *themselves* about how the sentences should be categorised at each time point, and were compared with intra-participant agreement values, to assess how different these two sets of values were. A t-test showed that participants had significantly better agreement with themselves than with the other participants in the task, and when inter- and intra-participant agreement values were ranked, intra-participant values typically ranked in at least the 75^{th} percentile. It was predicted that

if word senses do have some form of representation in memory, then participants should agree with themselves better than they do with others. Both forms of analyses indicate that this is the case.

While intra-participant agreement was typically found to be better than inter-participant agreement, this was not always the case. Indeed, some participants agreed with themselves about how the sentences should be sorted less than would be expected by chance. On the other hand, some participants' sorting decisions did not change at all. In sum, we observed variation in intra-participant agreement values that are difficult to explain in terms of an account in which word senses have fixed representations, à la Tyler and Evans (2001). Both this finding, of variation in intra-participant agreement, and the significant tendency for participants to agree with themselves better than with other participants, can be accommodated if we understand word senses in exemplar-theoretic terms. Under this account, exemplars – in this case, individual sentences – are stored in memory. Their membership of a particular category, however, is not. Instead, categories are constructed on the fly in response to the demands of a categorisation task. In principle, given that the task instructions do not change, the demands of the task, which should determine the strategy the individual uses to categorise the stimuli, should not differ at each time point. However, external factors such as a difference in time constraints and differing attention to characteristics of the sentences might result in participants' sorting strategies changing. In an exemplar model, this accounts for why some participants reached poor consensus with themselves between T1 and T2.

This interpretation was checked by testing a central prediction of the exemplar theory of categorisation: that categorisation is task-specific and that the relative similarity of stimuli is a product of the criterion/criteria the individual selectively attends to when categorising them. This was tested by manipulating whether participants sorted stimuli representing a single type of sense – either spatial or non-spatial – or a mix of both types of sense. I predicted that if word sense storage can be modelled in terms of the exemplar model of categorisation, stimuli that appear in both types of conditions would be categorised into more groups in the single sense-type condition than the mixed sense-type condition. The number of groups participants used to categorise these recurring stimuli was measured and an

independent samples t-test was used to establish whether a significant difference was found. A highly significant difference in the predicted direction was found.

Finally, inter-participant agreement values were calculated to establish how well participants agreed with each other over how the stimuli should be sorted. Across all 24 tasks, a wide range of agreement both within and across tasks was observed. If we understand the groups participants create to represent word senses they find meaningful, this finding indicates that individual speakers do not necessarily share word senses. This finding is compatible with the exemplar model.

4.7.1 An exemplar model of word sense representation

This study aimed to study the mental representation of word senses by asking whether they are stored in memory. Since some form of non-temporary representation was indicated, a cognitively-realistic account of their storage was also pursued. Cognitive linguists adhere to a theoretical framework that emphasises the necessity of explaining linguistic phenomena in a manner that is compatible with what we know about cognition more generally. Cognitive linguistic literature on the representation of word senses in memory has traditionally argued in favour of their representation in a radial network model (e.g., Brugman and Lakoff, 2006 [1988]; Tyler and Evans, 2001), in a manner that draws on and is broadly consistent with the prototype model of categorisation developed by Rosch and her colleagues during the 1970s (Rosch and Mervis, 1975; Rosch et al., 1976; Rosch, 1978). More recently, however, cognitive linguists (e.g., Divjak and Arppe, 2013; Gries, 2015) have considered whether an alternative model of word sense representation corresponds to empirical observations, specifically, a linguistic application of the (generalised) context model proposed by Medin and Schaffer in the late 1970s, and developed by these authors and Nosofsky from the 1980s onwards.

Based on the data presented in this chapter, I argue that word senses might be best understood as potential categories of stored exemplars of polysemous words that are constructed as and when needed according to a relevant classification criterion. Under this account, senses consist of exemplars of a given polysemous word that are considered by the individual to have equivalence in meaning *at the moment of disambiguation*. This entails that word senses may, therefore, not be fixed

representations. Nonetheless, the potential for the categorisation of a set of exemplars into identical groups on several occasions is not ruled out, as long as the criterion/criteria for judging similarity and therefore constructing the category remains constant. I therefore argue for an exemplar model of word sense representation.

The indication that word senses are *potentials* and do not have a fixed representation is not evidence in support of the monosemy position, which proposes that words have abstract, unitary meanings that are disambiguated on the fly (Ruhl, 1989). The finding that participants typically reach a level of agreement with themselves that is well above chance, and that they agree with themselves significantly better and more often than they do with other participants, is difficult to explain if we understand individual exemplars of polysemous words as disambiguated tokens which are discarded after they have served their communicative purpose.

On the other hand, this proposal that word senses may be mere potentials, rather than having a fixed and stable representation, presents a challenge to existing models of word sense representation which draw on the prototype model. For instance, Tyler and Evans (2001) argue that word senses are "distinct meanings instantiated [by which they mean stored] in memory" (p. 746), and propose a distinction between senses of the polysemous word *over* that are stored in memory, and meanings that are "interpretations produced on-line" (p. 727) and which presumably do not have the fixed representational status that word senses do. The evidence gathered here is at odds with this aspect of Tyler and Evans' theoretical model of word senses.

Further, the data reported here are difficult to reconcile with a representation system based on the prototype model of categorisation. Unlike the exemplar model, the prototype model does not predict that categorisation decisions are determined by selective attention. The effects of selective attention were observed in this study: participants in the single sense-type condition appeared to categorise the stimuli systematically differently to how those in the mixed sense-type condition did. A prototype-based account of word sense representation cannot account for these effects, nor for selective attention more generally. While work grounded in prototype theory by Labov (1978) has proposed that categorisation decisions can be modulated

by context, Labov's proposal called for variable feature weighting, a requirement better aligned with exemplar theory, which specifically predicts that features have variable weights, than with prototype theory, which does not predict that weighting may vary. Selective attention dictates that categories – including linguistic categories – are potentials, constructed in response to a particular (set of) categorisation criterion/criteria. Accordingly, individual exemplars of a polysemous word can belong to a number of categories depending upon (a) who is doing the categorising, and (b) what characteristics the individual selectively attends to when distinguishing between exemplars. Gahl and Yu (2006) claim that "each exemplar may belong to many categories simultaneously" (p. 213). This claim entails that an individual exemplar is located in multiple positions in the same multidimensional space simultaneously. This is an interesting claim but one which lacks empirical support, and which would presumably depend on a potentially infinite number of multidimensional spaces containing copies of all exemplars. I counter that while simultaneous membership of multiple categories is not possible under an exemplar model, *sequential* variation in category membership is not only a more realistic claim, but also one which follows logically from Medin and Schaffer's and Nosofsky's claim of the role of selective attention in categorisation. It seems difficult to imagine a prototype-based account of word sense representation in which individual exemplars may belong to multiple categories because the prototype model claims that it allows for optimal cognitive economy without compromising informativeness. If an individual exemplar could belong to multiple categories, as was observed in this study, in order to account for the findings under a prototype-based representation system one would need to posit a vast – potentially infinite – number of categories and prototypes. While the prototype model does not call for minimal storage, permitting a potentially infinite number of categories pushes storage demands to the opposite extreme. In this way, the balance between cognitive economy and informativeness, which the prototype model seeks to achieve, would be compromised.

4.7.2 Theoretical and practical implications

The findings in this chapter have certain theoretical and practical implications. They also raise a number of questions that might be the topic of further research. These are discussed below.

4.7.2.1 Theoretical implications

First and foremost, the findings of this study support an exemplar model of word sense representation. While an exemplar-theoretic account of word senses was hinted when discussing the finding that individuals agree with themselves better than they do with others about how stimuli should be sorted, and that there appear to be individual differences in word senses, the most significant support for this account came by testing the prediction that categorisation decisions are shaped by selective attention. This prediction was substantiated.

By testing this particular aspect of the exemplar model of categorisation, this study extends exemplar-theoretic accounts of the storage of linguistic information. While there is a growing body of research examining the role of frequency and repetition (e.g., Bybee, 2002, 2006; Divjak and Arppe, 2013; Pierrehumbert, 2000), to my knowledge there is little research in linguistics generally, and none in semantics specifically, that addresses the central claim of Medin and Schaffer's (1978) and Nosofsky's (1986) exemplar model: that categorisation decisions are determined by selective attention. Finding support for this claim has weakened the case for a prototype-based model of word sense representation, and entails that word senses might not be understood as either fleeting entities constructed and then forgotten, nor as fixed representations, but might be better understood as something in between. I argue that an exemplar-theoretic account of word senses views senses as nothing more than potential categories comprising previously encountered exemplars, which *may* be recreated, but which do not have a fixed representational status. This finding offers a theoretical explanation and support for Kilgarriff's claim in the lexicographic literature that "word senses exist only relative to a task." (Kilgarriff, 1997, p. 1).

4.7.2.2 Practical implications

4.7.2.2.1 The status of expert intuitions

Individuals appear to classify exemplars of polysemous words in different ways, despite receiving identical guidelines instructing them to categorise the stimuli *only* on the basis of the meaning of the target polysemous word. This presents a challenge to the intuition-based study of polysemy at a theoretical level, which has traditionally been the focus of research in cognitive linguistics, and at a practical level, which is

the focus of research in word sense disambiguation in computational linguistics. Put simply, these findings discourage an intuition-based analysis of polysemous words, and authors should explicitly acknowledge that the distinctions that they find meaningful might not be meaningful to other speakers. In the face of the findings presented here, I would argue that an intuition-based analysis of the senses of polysemous words is unlikely to prove generalisable or cognitively realistic. I recommend, on this basis, that analyses of the senses of polysemous words are informed by empirical evidence. I also respond to Talmy's claim that "introspection has the advantage over other methodologies in seemingly being the only one able to access [meaning] directly." (2007, p. xiii). The fact that participants in this set of studies have produced groups of sentences that appear to follow some logical and non-random categorisation principles suggests that Talmy may indeed be correct here. However, his claim, which forms part of his larger argument over the utility of introspection and intuitions in linguistic research, does not account for the possibility – which has been realised in this study – that when an individual encounters a pair of examples of a polysemous word, the meaning(s) they access to interpret them may differ to those which another individual accesses. This finding entails that, while introspection may be a useful tool for understanding word meaning, meanings identified by introspection are subject to variation across individuals. Consequently, the use of introspection alone as a means of studying word meaning may produce results that do not faithfully represent meanings held by other speakers. This finding is compatible with and adds to existing literature that problematises the utility of expert intuitions in the description of linguistic – primarily syntactic – phenomena (Bradac et al., 1980; Dąbrowska, 2010; Gibbs, 2006; Gordon and Hendrick, 1997; Labov, 1972; Miller, 1962; Ross, 1979; Schütze, 1996; Schwarz-Friesel, 2012; Spencer, 1973).

4.7.2.2.2 Automatic word sense disambiguation

Automatic word sense disambiguation is a necessary component of a successful artificial intelligence system. Human intuitions about word senses are harnessed to train an automated WSD algorithm, and to assess the disambiguation results. The development of an inventory of word senses has recently seen significant methodological development in the changing use of expert and naïve intuitions about where word sense boundaries lie. Specifically, word sense disambiguation research

increasingly recruits the "crowd", groups of anonymous and naïve participants, to categorise exemplars of a given word to one or more predefined semantic categories. The findings of this study raise questions over the utility not only of expert intuitions, but the role of naïve intuitions in the development of a word sense inventory. If individuals have different senses of a given word – and if their decisions about what category an individual exemplar belongs to is not permanent, but subject to change depending on what the category options are – then how reliable are the decisions they make if the goal is to develop a "gold standard"?

Based on the findings presented here, I argue that fine-grained decisions are not a particularly useful dataset, and that a fine-grained approach to word senses may never be fruitful, no matter how much training the algorithm receives, nor how many human participants (expert or otherwise) contribute to the training set and sense inventory. I would not, however, argue that good automated WSD algorithms based on an inventory compiled by human informants are impossible. Instead, I recommend that coarse-grained senses are more useful. I reach this conclusion based on the conflicting realities of both individual differences in word senses, as observed here, and nonetheless effective communication between individuals. I therefore offer an evidence-based argument compatible with Ide and Wilks' (2007) recommendation that computational WSD tasks do not use "the standard fine-grained division of senses", but focus instead on "broad discriminations" (p. 47). I explore this in more detail in the following section.

4.7.3 Questions raised

4.7.3.1 Successful communication in the face of individual differences in word senses

If individual speakers' senses of a given word do not overlap, how is successful communication achieved[9]? At face value it may seem problematic to posit an account of major individual differences in word senses. After all, if the meaning that a pair of individuals attributes to a single exemplar of a polysemous word differs, surely the meaning intended by the speaker will not map onto the listener's interpretation. However, successful communication takes place between individuals with markedly different linguistic systems (e.g., children and their caregivers, L1 and

[9] As Ferreira et al. (2002) note, however, communication is not always successful.

L2 speakers), and when speech is disfluent (Ferreira et al. 2002). In a study that compared computational models' and native speakers' performance in their choice of six synonymous Russian *try*-verbs, Divjak et al. (2016, p. 27) proposed that multiple computational models of grammar can correspond to human usage. Indeed, they recommend that the pursuit of a "single 'best' model" of human grammar should be side-lined in favour of developing multiple models, representing individual variation in grammar. In a somewhat different explanation of successful communication in the face of individual differences in grammar, Ferreira et al. (2002) propose that successful sentence interpretation in normal communicative situations is based on a "good enough" model of sentence processing. Further, successful processing and interpretation is supported by contextual information. We might extend this argument to account for the incongruity between variation in word senses and (generally) successful communication. In the study reported here, participants were engaging in a disambiguation experiment rather than in a conversation that required disambiguation. This artificial scenario thus removes contextual information that a conversation in a might provide. In a natural communicative situation, under an exemplar model of word senses, it might be the case that context provides information and disambiguation/categorisation biases that serve to shrink or extend dimensions, thus resulting in different categorisation decisions than the ones observed here.

This account assumes that successful disambiguation depends on finely-grained word senses. Computational linguists have claimed that, on the contrary, disambiguation beyond even the homograph-like level is rarely necessary for human and computer understanding, and that division of these very coarse senses into more finely-grained sub-senses is only done if successful communication depends on it (Ide and Wilks, 2007, p. 66). This is not incompatible with Ferreira et al.'s argument. Let us consider Ide and Wilks's very coarse-grained senses, which have the potential to consist of more finely-grained senses. It seems feasible that the senses that emerge in the sorting task reported here might, at least by some participants, be collapsed into a smaller number of coarser senses. If we were to pair up two such participants in a conversational scenario, it is possible that these coarser senses might overlap. Let us take as an example the categorisation decisions made by participants BMi15 and BMi16, in the first mixed sense-type task for *below*, looking specifically at the

category to which they assigned the sentence *The sleeves gradually get tighter and end below the elbow*. Participant BMi15 assigned it to a single-member group labelled 'beyond, on a dimension which is implicitly downward'. Participant BMi16 assigned that sentence to a group labelled 'beyond', along with three other sentences. The categorisation decisions, when adopting a fine-grained approach, therefore differ. However, it might be the case that both participants would agree that this sentence is an exemplar of a broader INFERIOR POSITION sense. Indeed, in the network visualisation that was produced using data from this task, this sentence was located in a seemingly catch-all, generic spatial cluster. If this sentence arose in conversation between BMi15 and BMi16, and if the success of the conversation depends on accessing a finely-grained sense, a mismatch in interpretation might occur. However, if accessing a broad sense allows for "good enough" interpretation, no mismatch will occur, since they both belong in the same broader, collapsed sense.

As Divjak et al. (2016) observe, multiple computational models of grammar correspond to actual human behaviour. At face value, multiple models of word sense categories might be needed to explain the individual differences observed in this study. However, in the face of evidence which supports an exemplar-theoretic model of word sense representation, I would argue that no fixed model of word senses – multiple or otherwise – is cognitively real, given that actual disambiguation decisions are subject to not only external variation, but internal variation, too.

4.7.3.2 Why do some participants agree with each other more than others?

As noted previously, this study has shown that some pairs of participants reach a high – sometimes perfect – degree of consensus over how the stimuli should be categorised. The opposite extreme was observed too, with some pairs of participants reaching below-chance agreement. At this point, it is not clear what factors might explain these observations. Infrequent exposure to the words tested here might result in fewer opportunities to identify the characteristics of its meaning which are most important in establishing its meaning in context. For example, limited exposure to *above* in written contexts may result in a different categorisation decision for the exemplar *It was refused for the above reasons* to a participant with more exposure to written language. It is beyond the scope of this thesis to explore the reasons for differences in agreement across participants, but collection of demographic data

makes a study of the role of education and type of work in inter-participant (dis)agreement possible in future research.

4.7.3.3 Why do some participants agree with themselves more than with others?

Just as reasons for differences in inter-participant agreement are unclear at this point, so too are the reasons why some participants reached perfect consensus with themselves, and others agreed with themselves less than would be expected by chance. An exemplar model of word sense representation predicts that differences in sorting decisions are attributable to differences in sorting strategies; a participant may have had more time available to complete one of the tasks than the other, which might have changed how much detail they attended to in the meaning distinctions. Equally, a participant may have interpreted the instructions differently on each occasion. As was the case in observations of inter-participant agreement, intra-participant (dis)agreement may also be determined by differences in education, and type of employment.

4.7.3.4 What can networks tell us about word meaning?

Finally, the study has raised questions over the utility of network visualisations of sentence-sorting data in the study of word senses. When agreement amongst participants over how sentences should be categorised is high, high quality networks can be produced. The communities that emerge in these networks might be tentatively taken as representing word senses. The fact that the communities detected in these high quality networks have comprehensible and seemingly non-random membership encourages further study of the role they might play in understanding word senses and their interrelations. They may prove useful, for example, in identifying whether finely-grained senses can be collapsed into the larger, homograph-like sense that Ide and Wilks (2007) claim to generally be the most useful level of meaning for successful disambiguation. Further, the fact that I found a strong and significant correlation between mean inter-participant agreement values for each task, and the modularity value (which measures the quality of the network) of the resultant network gives further weight to using them in future research.

4.8 Conclusions

Using a set of open sentence-sorting tasks with a large sample of participants, and by performing statistical analyses on the resultant data, a significant body of evidence was gathered which collectively indicates that word senses have some form of

mental representation, and might be best understood as potential categories of stored exemplars of polysemous words. In brief:

1. Participants agreed with themselves better than with other participants over how stimuli should be sorted. This was taken as evidence that word senses do have some form of mental representation.
2. Sorting decisions differed according to the number of broad semantic categories represented by the sentences in the stimuli sets. This was taken as evidence that categorisation may be task-based and that decisions about the similarity of exemplar sentences may be affected by the characteristics of the meaning of the target word the individual selectively attends to.
3. Individual differences in how sentences should be sorted were observed. On the basis that the groups participants created were taken as indicators of word senses, I argue that individuals may have different senses of polysemous words.

Modelling linguistic phenomena in terms of the exemplar model is not a new approach in (cognitive) linguistics; for example, exemplar theory has been invoked to account for phonetic variation (Bybee 2002; Pierrehumbert 2000) and grammar (Bybee 2006). Equally, it has been recommended as a suitable means of modelling semantic phenomena, including polysemy (Gries 2015) and, in conjunction with varying abstraction models and prototype representations, near-synonymy (Divjak and Arppe, 2013). However, the findings made in this study build upon this literature by testing a claim central to the theory: that categorisation is task-based and that the category a particular exemplar is assigned to is a factor of the criterion/criteria that the categoriser selectively attends to. In this way, this study extends the literature on exemplar-theoretic accounts of linguistic categorisation and representation. This aspect of the study also provides support for, and a theoretical justification of Kilgarriff's claim in the lexicographic literature that "word senses exist only relative to a task." (1997, p.1).

Chapter 5 Discussion and conclusions

Polysemy, the phenomenon whereby a word has a number of distinct but (arguably) related senses, is an issue that concerns a great number of researchers working both within and beyond the cognitive linguistics community. Indeed, scholars in the field have claimed that the study of polysemy is "rampant" (Cuyckens and Zawada, 2001, p. xv). However, there remains more to be said. This thesis aims to add to, and develop, our understanding of polysemy, and in particular, the senses of polysemous words. While polysemy is not a topic that is of unique interest to cognitive linguists, what distinguishes polysemy research in cognitive linguistics from work on this topic by scholars working within other theoretical frameworks is that an account of the *psychological status* of polysemous words is pursued. The motivation for this pursuit is cognitive linguists' adherence to the cognitive commitment, i.e., that claims we make, and theories we develop, about the nature of language must be compatible with what we know about the brain and mind more generally.

Polysemy, which entails a potential proliferation of individual word senses, is therefore an account of word meaning that is in conflict with monosemy. In contrast to polysemy, proponents (e.g., Ruhl, 1989) of the monosemy approach claim that words are ambiguous, with a unitary, underspecified meaning which is substantiated on an ad hoc basis using surrounding context, such as sentential and environmental context. Word meaning is therefore not, under a monosemy account, stored in memory. Theorists who support the polysemy approach claim that context does play a role in deciphering word meaning, but that context itself does not create this meaning. For example, Gibbs and Matlock (2001) propose that context *facilitates* the discrimination of word senses.

5.1 Expert intuitions, and individual differences

Amongst the vast literature on polysemy exist a number of analyses of polysemous words. The studies by Brugman and Lakoff (2006 [1988]) and Tyler and Evans (2001), both of which focus on *over*, are canonical examples of efforts to pin down the senses of this polyseme, and offer an account of the relationships between senses. These analyses are made on the basis of the authors' intuitions. Certainly, Tyler and

Evans describe their route to isolating word senses as principled, and aim to offer a constrained set of senses. Of course, the role of intuition in the practice of identifying word meaning has a strong precedent; in his "idealized" account of how lexicographers identify word senses for publication in dictionaries, Kilgarriff (2007, p. 31) describes a highly intuition-led procedure. However, in the face of criticism, levelled in particular at syntacticians, over the status of intuition-based studies of linguistic phenomena, this approach is problematic. Moreover, there are growing calls amongst cognitive linguists to adopt empirical approaches to linguistic scholarship (e.g., Arppe and Järvikivi, 2007). For these reasons, the assumptions inherent in intuition-led analyses of polysemous words, namely that expert intuitions correspond to naïve speakers' intuitions, and that word senses are shared by native speakers of a given language, require testing. This brings me to the first aim of the thesis: to study whether the sense distinctions that I, as an expert, are shared with naïve and expert participants who speak English as their native language. The second assumption, that word senses are shared by native speakers, is not unreasonable. If we didn't agree on what the senses of a given word are, how would communication be successful? However, if we observe a non-native and native speaker of English in conversation, or witness communication between a parent and toddler, we will observe that in spite of the considerably different linguistic systems held by each speaker, communication can be successful. Moreover, individual differences have been observed in other areas of language, such as grammatical attainment (e.g., Street and Dąbrowska, 2010), language acquisition (e.g., Bates, Dale, and Thal, 1995) and metaphor interpretation (Duffy 2015). Individual differences in word senses need not, therefore, present major problems to successful communication, and they would be compatible with what is already known about language. For this reason, the second aim of this thesis was to assess whether there is evidence that individuals do not reliably agree on the senses of polysemous words.

5.2 Mental representations of word senses

Certain polysemy theorists in cognitive linguistics propose that word senses have some form of fixed mental representation. In their account of the representation of the senses of the word *over*, Tyler and Evans (2001) argue that some – but not all – senses of polysemous words are stored in memory. They go on to propose that stored word senses are organised in a radial system, and bear some relation with a

prototypical sense. In this way, Tyler and Evans' proposal borrows from the prototype theory of categorisation developed by Rosch and her colleagues during the 1970s (Rosch and Mervis, 1975; Rosch et al., 1976; Rosch, 1978), in a way that is not uncommon in cognitive linguistics (for examples of linguistic applications of the principles of prototype theory, see Coleman and Kay, 1981; Evans and Tyler, 2005; Gilquin and McMichael, 2008; Ibarretxe-Antuñano, 2004; Ibbotson, Theakston, Lieven, and Tomasello, 2012; Ibbotson and Tomasello, 2009; Langacker, 1986; MacLaury, 1989, 1991; Rice, 1996). Certainly, elsewhere in cognitive linguistics scholars have argued that polysemous words are examples of categories (e.g. Taylor, 2003). This motivates an account of the representation of word senses that is compatible with general models of categorisation.

More recently, Divjak and Arppe (2013) have considered the representation of other semantic phenomena, this time near-synonymous Russian verbs of *trying* and *thinking*. Contrary to earlier work, which has favoured a prototype-based account of meaning representation, Divjak and Arppe conclude that a single representational model cannot account for their observations. Instead, they propose that meaning may be modelled in terms of a number of different types of representations, from full prototype to individual exemplars, and at different degrees of abstraction in between these two polar positions. Elsewhere, Gries (2015, p. 482) has described word senses in a manner akin to an exemplar-based model, using terms such as "multidimensional semantic space", regions of which denote individual senses. Indeed, his collaborative work on behavioral profiles is compatible with an exemplar model (e.g., Gries and Divjak, 2009). Behavioral profiles describe a vast number of characteristics of an exemplar of a given word. These characteristics, when understood in exemplar-theoretic terms, correspond to the dimensions that shrink and expand in a categorisation task. A further example of an exemplar-oriented approach to word senses is Murphy's (2007) "limited listing" approach. In his account, Murphy proposes that encounters with a polysemous word are mapped to a position in a multidimensional space. Gries (2015) notes that there is a growing consensus that word senses are represented in an exemplar-based form. This shift in focus away from prototype-oriented accounts of the representation of words senses, towards an exemplar-theoretic account, motivates the present study.

The exemplar model of categorisation (also known as the [generalized] context theory of classification learning [Medin and Schaffer, 1978; Nosofsky, 1986]) predicts that stimuli have perceptual features that can be modelled in a multidimensional space. Exemplars of a given concept, such as exemplars of a DOG, are tagged for relevant perceptual characteristics, such as colour, amount of fur, type of bark, and so on. Dimensions in the multidimensional space represent these characteristics, and the value of a given dimension for example, brown fur, is plotted. Exemplars that have similar values on a given set of dimensions are positioned more closely together than those that have different values on that set of dimensions. As a result, they are judged to be more similar. Of crucial importance in this theory is the argument that these dimensions are *dynamic*; they can shrink or extend depending on whether or not the characteristics they correspond to are the object of the *selective attention* of the categoriser. When a characteristic is selectively attended to, the associated dimension expands, thus revealing fine distinctions along it. Conversely, when a characteristic is irrelevant, its dimension shrinks, thereby hiding distinctions.

The proposal that word senses are stored in memory is not a new one. Neither is the suggestion that they have an exemplar-based representation. However, to my knowledge, no research has tested this central prediction of the exemplar-based theory of categorisation – that categorisation judgments are determined by selective attention – in a study of polysemous words, or indeed any semantic phenomenon. While other linguists have predicted – and, where they have tested for them, found – selective attention effects in linguistic categorisation (Ellis, 2006; Francis and Nusbaum, 2002; Kalyan, 2012; Lively et al., 1993), others (e.g., Bybee, 2006), omit selective attention from accounts of linguistic categorisation, and therefore treat linguistic categories separately from categories of other phenomena. This brings me to the final two aims of this thesis. The research presented here empirically tests (1) whether there is evidence that word senses are stored in memory and, if they are, (2) whether they are represented in a manner consistent with what is already known about the exemplar theory of categorisation. To answer that, I ask whether linguistic categories – specifically, word senses – are subject to the selective attention effects that are a central component of the exemplar model.

5.3 Chapter structure

This chapter draws the thesis to a close. It summarises the purpose and primary findings of each experiment, each of which concentrates on the polysemous words *over, under, above* and *below*, and discusses what these findings add to existing knowledge. It then discusses the collective implications of the findings of the three sets of experiments. It closes with some concluding remarks on the place of the findings reported here in the broader research landscape, and makes recommendations for further work, in light of questions raised during the course of the research.

5.4 Chapter 2. Experiment 1: A closed sentence-sorting study to test the representativity of linguists' intuitions about word senses

Polysemy is a hot topic in cognitive linguistics, and a vast literature on polysemy exists. Brugman's master's thesis (1981), along with her later work with Lakoff (1988) initiated a body of research that examined the organisation of the senses of polysemous words, relative to each other, and relative to a prototypical sense. Briefly, they argue that polysemous words are radial categories, with senses connecting to a prototypical sense, or an intermediary sense. Tyler and Evans' (2001) paper, which also studied *over*, has clear ties with Brugman and Lakoff's original work. While in this seminal work they do not specifically describe the structure of the senses of polysemous words as radial, the diagrammatic representation of the senses of *over* (p. 746), is clearly inspired by the radial structure proposed by Brugman and Lakoff. In later, related work on the preposition *in*, Evans and Tyler (2004) confirm that they do indeed model polysemous words as radial categories, after Brugman and Lakoff. Tyler and Evan's case study of *over* was developed in part in response to doubts raised about the cognitive reality of existing accounts of the representation of polysemous words. They state, for example, that Sandra and Rice (1995) claim that analyses of polysemous words were theretofore arbitrary and influenced by the analyst's "preferences (or indeed imagination)" (p. 733). In response, Tyler and Evans develop a principled approach to the isolation of word senses and identification of the prototypical sense, represented by a "protoscene" (p. 735), and make claims as to the cognitive reality of the senses they identify, claiming that some are stored in long-term memory.

While Tyler and Evans' efforts to develop a principled approach to isolating word senses and a protosense are valiant and represent a clear response to Sandra and Rice's (1995) criticism over the role of the analyst's intuitions, they are inherently flawed in that the approach they advocate also depends on intuitions and introspection. Certainly, in their methodology for identifying the primary sense, one aspect is objective: they state that an indicator of the primary sense is that which is attested earliest (p. 734). However, the other means by which they identify the primary senses, namely predominance in a network, relations to other prepositions, and grammatical predictions, is open to author bias in two ways. First, the *author* judges the predominance in the network, relations to other prepositions, and grammatical predictions, based on their intuitions. Second, the *author* decides what the senses are in the first place. Certainly, these senses are, according to Tyler and Evans, arrived at in a principled manner. Principled as that might be, their identification proceeds solely on the basis of intuitions.

In light of existing research that questions the reality and utility of linguists' intuitions in the study of other linguistic phenomena (Bradac et al., 1980; Dąbrowska, 2010; Gibbs, 2006; Gordon and Hendrick, 1997; Labov, 1972; Miller, 1962; Ross, 1979; Schütze, 1996; Schwarz-Friesel, 2012; Spencer, 1973), I was unconvinced by the reality of the senses that Tyler and Evans propose. In response, and acknowledging the status of polysemous words as linguistic categories, I set out to determine whether, in a categorisation task analogous to those used in work by Cuyckens et al. (1997) and in computational linguistics (Bhala and Abirami, 2014), naïve and linguist participants would agree with the senses that *I*, as a trained linguist with particular expertise about the meanings of the words *over, under, above* and *below*, found meaningful. The results of the categorisation tasks, operationalized as sentence-sorting tasks, indicate that my intuitions about what the senses of four polysemous words are do not consistently and reliably correspond to those held by other native speakers of English, regardless of whether or not they too are linguistics experts. Two explanations of this outcome were identified. On the one hand, the participants in the study may have agreed what the senses of these four words are, and it is simply the case that these senses do not correspond to mine. On the other hand, individuals may have different senses of a given polysemous word.

An important outcome was that there were differences in the degree to which individual participants and I agreed about how the stimuli should be sorted. Variation in agreement between me and each participant problematises, but does not rule out, the interpretation that the participants had a homogenous set of senses they found meaningful within the stimuli, and that this set contrasted with mine. If participants agree about what the senses of a word are, they *should* (dis)agree with my sense distinctions to the same extent. This was not the case. This outcome could, however, be explained on the basis that the task was highly structured, and participants may have used this structure to infer the "correct" sorting solution. In this case, rather than categorising the stimuli in a way that corresponded to their intuitions about word meaning, participants were instead attempting to figure out what the final solution *should* look like, assuming that there was a "correct" outcome. Certainly, clues as to a possible solution were available: they may have correctly guessed that I judged each category to have six members, and the distinctions evident in the different group labels may have hinted at distinctions within the stimuli.

5.5 Chapter 3. Experiment 2: An open sentence-sorting task to test for individual differences in word senses

In light of the indication that participants agreed with me to different extents about what the senses of the words *over, under, above* and *below* are, an open sort task was devised to explicitly address the possibility that there are individual differences in word senses.

Individual differences have been observed in a range of aspects of language acquisition (Bates, Dale and Thal, 1995), including early phonology (Leonard, 1980), joint attention skills (Mundy and Gomes, 1998), lexical processing (Fernald and Marchman, 2012), the development of multiword speech (Pine and Lieven 1993; Shore, 1995), vocabulary (Bates et al, 1995), syntax (Vasilyeve et al., 2008) and semantics (Rice, 2003). Individual differences in adult language are the subject of growing attention. For example, Street and Dąbrowska have found evidence of individual differences in grammatical attainment and processing that appear to be based on differences in linguistic experience (Dąbrowska and Street, 2006; Street and Dąbrowska, 2010, 2014). Divjak et al. (2016, p. 27) have recently gathered data

suggesting that "a very large number of models" of grammar correspond to attested usage, thus providing further evidence that there are individual differences in grammar. In contrast, I am aware of no literature that directly investigates whether or not there are individual differences in word senses. Certainly, data from research in computational linguistics suggests that there *may* be individual differences. For example, word sense disambiguation research has shown that individuals "sense-tag" examples of polysemous words differently (e.g., Passonneau, Bhardwaj, Salleb-Aouissi, and Ide, 2012). However, these findings were gathered in a manner analogous to the closed sorting task described above, meaning that they cannot conclusively establish that different people have different senses of a given word, since it may be the case instead that a sense inventory developed by an expert is simply different from a single set of senses that all other speakers find meaningful.

I investigated whether or not there may be differences in word senses across individuals by asking naïve participants to complete a large-scale open-sort task. In this case, participants were required to create their own categories that captured examples of the target word with the same meaning. Inter-participant agreement was calculated with Morey and Agresti's adjusted Rand, which returns a value ranging from -1, representing total disagreement, through 0 (chance agreement), to 1, representing perfect agreement. Using Neuendorf's (2002, p. 3) recommended 0.8 cut-off point for acceptable agreement, the results suggest that, on the whole, participants fail to agree with each other about how examples of a polysemous word should be categorised. If we understand the groups participants create to represent the senses they find meaningful, this indicates individual differences in word senses. Accordingly, the findings of the first, closed-sort experiment are explained. Participants failed to agree with me not because my word senses differed to participants' word senses, and participants themselves agreed with each other about what the senses of these four words are. Instead, it seems more likely that they disagreed with me because individuals have different senses of polysemous words.

I noted in that chapter that methodological concerns compromise the surety of this conclusion. The scale of the task was vast; participants were required to sort 100 sentences that were unedited to make them well-formed. Participants therefore needed to read each sentence and make a judgment as to what the meaning of the

target word was. They then had to match up sentences which used the target word in the same sense. This is a difficult task, requiring participants to remember a potentially large number of senses and judge which sense a particular exemplar sentence corresponded to. The scale of this task may therefore have caused fatigue, boredom and semantic satiation, which may have resulted in categorisation decisions that lacked coherence and consistency.

5.6 Chapter 4. Experiment 3: Testing an exemplar model of word senses

The sentence-sorting task is highly versatile and this versatility was exploited to answer other interesting questions about word senses. Given the large scope of this series of this study, the following section will be divided into smaller subsections addressing (1) individual differences in word senses, (2) storage of word senses in memory, and (3) selective attention in linguistic categorisation judgments. The section will end with a brief conclusion summarizing the findings.

5.6.1 Individual differences in word senses

The possibility raised in Experiment 2 that individual speakers have different senses of polysemous words is an exciting one. It is both under-investigated and compatible with what we already know about individual differences in other aspects of language. However, methodological factors may explain the finding of poor inter-participant agreement. To address this possibility, I devised a smaller sentence-sorting task. Like the task just discussed, this too was an open-sort task, requiring participants to categorise sentences into groups of their own making. This time, however, the task was much smaller in scale, with just 36 sentences. Further, the stimuli were structured in that I judged them to represent six examples of six senses. These factors should limit fatigue and therefore produce more consistent and coherent sorting decisions. Inter-participant agreement values were highly varied across all 24 sorting tasks, reflecting varying degrees of (dis)agreement over how stimuli should be sorted. If we understand the categories of sentences produced by participants to reflect the senses they find meaningful, this outcome suggests that individuals do not necessarily share word senses.

5.6.2 Storage of word senses in memory

The task was also used to address an important theoretical issue that has been the focus of cognitive linguistic treatments of polysemy: the representation of word

senses. Accounts of the storage of, and relationships between word senses have traditionally invoked a prototype-based system, thus borrowing from Rosch and colleagues' prototype model of categorisation (Rosch and Mervis, 1975; Rosch et al., 1976; Rosch, 1978). While the prototype model produced powerful explanations of observations such as typicality effects in categorisation decisions, other models have been proposed which can also account for typicality effects. Indeed, other models, specifically the exemplar model developed by Medin and Schaffer (1978) and Nosofsky (1986) better account for observations made in categorisation experiments than the prototype model (Medin and Schaffer 1978; Murphy 2004, p. 103). Indeed, and while prototype-based approaches are favoured in the field, (cognitive) linguists have considered exemplar-theoretic models of linguistic representation to a limited extent (Chandler, 2015, p. 3). For example, the exemplar model has been adopted to account for other linguistic phenomena, such as constructions (Bybee, 2006) and phonology (Pierrehumbert 2000). However, canonical research in cognitive linguistics has assumed a prototype-based representation of polysemous words and their senses, and has argued that word senses are stored in long term memory (e.g., Brugman and Lakoff, 2006 [1988]; Tyler and Evans, 2001).

Experiment 3 aimed to contribute to this literature by testing whether word senses are stored in memory. This was achieved by asking participants to complete an identical sentence-sorting task twice, divided by a period of two months. If word senses do have some form of mental representation, I predicted that participants would agree with themselves about how the stimuli should be sorted *better* and *more often* than with other participants. Specifically, I predicted that a single participants' sorting decisions at T1 and T2 would show more agreement than between their sorting decisions at T1, and other participants' sorting decisions at T2. The data revealed that participants agreed with themselves significantly better and more often than with other participants. However, it was also observed that participants varied in how well they agreed with themselves; it was not always the case that they agreed with themselves better than with others, and in individual participants the opposite was true. The fact that some participants seemed to make very similar – if not identical – sorting decisions while others made very different decisions is at odds with the fixed representation that scholars such as Tyler and Evans (2001) claim

word senses to have, and is difficult to explain under the prototype model typically invoked. Equally, however, extremely high agreement within some participants is difficult to explain under a monosemy account which claims that word senses are created ad hoc and do not have any form of mental representation. An alternative account may therefore be necessary, and it was predicted that the exemplar model may be able to account for both very good and very poor intra-participant agreement.

5.6.3 Selective attention in linguistic categorisation decisions

In order to assess whether word senses can be understood as categories of exemplars in a manner consistent with the exemplar theory, the experiment also tested for evidence of one of the predictions of the theory: *selective attention*. Selective attention, as defined and discussed in Chapters 1 and 4, is a central component of the exemplar theory of categorisation developed by Medin and Schaffer (1978) and Nosofsky (1986). It is therefore surprising that it has received little attention in linguistic applications. A very small number of scholars have tested for and observed selective attention effects in linguistic categorisation (Ellis, 2006; Francis and Nusbaum, 2002; Lively et al., 1993), and computational models of inflectional morphology that do not incorporate feature weightings, and therefore selective attention, perform worse than those that do (Chandler 2010). Likewise, a recent theoretical account of differential acceptability of verbs in questions with long-distance dependences by Kalyan (2012) – a process that the author understands as a categorisation event – has reconciled seemingly opposing explanations of empirical findings by Ambridge and Goldberg (2008) and Dąbrowska (2004) by explaining the findings in terms of selective attention. However, by and large, linguistic applications of the exemplar theory have overlooked selective attention. Indeed, some literature proposes that linguistic categories are fixed entities, rather than dynamic groups of exemplars whose position in a multidimensional space – and therefore category membership – is modulated by selective attention (Bybee, 2006). In this way, it seems that some scholars advocate a more comprehensive application of the exemplar model to explaining linguistic phenomena, while others recommend incorporating only some aspects of the model.

In the interest of adhering to the cognitive commitment, I aimed to study whether a special case of the model, such as that described by Bybee, is necessary to explain

linguistic categorisation. Specifically, I tested the applicability of the exemplar model to this example of linguistic categorisation by searching for the effect of selective attention in participants' sorting decisions. This is a central aspect of both Medin and Schaffer's (1978) and Nosofsky's (1986) models, and one that Bybee proposes may not be present in linguistic categories due to their frequency. By testing whether selective attention was observable in this instance of linguistic categorisation, I could then answer the questions of whether word senses can be modelled as exemplar categories, and whether even highly frequent categories such as linguistic categories display selective attention effects, and are therefore dynamic and not the fixed entities that Bybee (and Tyler and Evans, 2001) appears to suggest that they are. I tested this by manipulating what type of stimuli participants were given. They were assigned to categorise one of three versions of stimuli: (1) a set of 36 examples of *spatial* uses of the target word, (2) a set of 36 examples of *non-spatial* examples, or (3) a set of both *spatial* and *non-spatial* examples. The first two versions of the task are labelled the "single sense-type" condition, and the third version the "mixed sense-type" condition. I predicted that the presence of both spatial *and* non-spatial meanings in the mixed sense-type condition would act as a highly salient semantic distinction within the stimuli. In line with the exemplar model, I then predicted that participants would selectively attend to this distinction. As a result, distinctions *within* these two broad groups, i.e., further distinctions within spatial examples, and non-spatial examples, would be less attended to, resulting in these distinctions being overlooked. I therefore predicted that stimuli that appeared in both the single and mixed sense-type conditions would be sorted into fewer, more coarse groups in the mixed sense-type condition, reflecting the fewer distinctions that participants would notice. This was shown to be the case.

5.6.4 Conclusions

The findings of this final study therefore provide evidence in support of the conclusion that word senses may be temporary and merely potential categories of exemplars of a polysemous word, and may be modelled in terms of the exemplar model developed by Medin and Schaffer (1978) and Nosofsky (1986). In this way, I move away from traditional linguistic categorisation accounts of polysemy, which propose that polysemous words are examples of linguistic categories (e.g., Tyler and Evans, 2001). The data gathered here suggest that senses *themselves* might be

(potential) categories. The study was not explicitly designed to test the possibility that polysemous words are not categories, and for that reason the status of words as categories cannot be conclusively ruled out. The data do not, however, provide support for the notion that polysemous words are categories. The study indicated that selective attention effects, predicted by the exemplar model to affect how individuals assign stimuli to a category, were observed at the level of senses. In other words, and continuing to make the assumption that the categories participants create reflect word senses, the senses they judged to be represented by the stimuli depended on what semantic features of the word in context they attended to. For example, in Experiment 3, participant BMi9 in the mixed sense-type condition for the *below* task uses two categories to classify a set of 24 sentences. They attend to a distinction between TEXT USES, and what they judge to be descriptions of 'physical' configurations. In contrast, participant BS10 in the single sense-type condition sorts the same sentences into six categories, therefore making much finer distinctions amongst the stimuli. In these examples, the participants are attending to different semantic features of the word in context, and it is these different features that license the creation of different categories. If it is polysemous words that are linguistic categories, that would entail that word senses are members of these categories. Consequently, we should not see selective attention effects at the level of word senses. On the contrary, we *did* observe this.

The findings also serve to develop the line of work on linguistic applications of the exemplar model. Bybee (2006) has applied an adjusted version of a traditional exemplar model to describe the representation of constructions. Chandler (2010) notes that research on morphology and phonology has successfully modelled these linguistic phenomena in terms of an exemplar model, and that other, smaller-scale research has studied the application of the model to syntactic phenomena. He asks whether the models he describes can "scale up to a more extensive model of language" (p. 412). The findings presented in this thesis suggest that the exemplar model developed by Medin and Schaffer (1978) and Nosofsky (1986) has application to semantic phenomena also, thus indicating that the exemplar model can indeed scale up to model language more extensively.

5.7 General discussion

As a whole, this thesis provides insights into particular aspects of the psychological status of the senses of polysemous words. Working within the cognitive linguistic framework, and making the theoretical and methodological assumption that word senses are linguistic categories, it set out to establish whether linguists' intuitions about word senses correspond to those of naïve speakers and other linguists. It then aimed to assess whether the findings could be explained in terms of what we already know about language – i.e., that it is subject to individual differences – and finally, it aimed to test whether word senses and the observations made in this thesis are compatible with an exemplar-theoretic model of categorisation. The findings provide empirical support for the conclusion that naïve speakers' and experts' intuitions about what the senses of a given word are fail to coincide, which is explained by the further finding that individuals appear to have different senses of polysemous words. The data also indicates that word senses have *some form* of representation in memory, and that they can be modelled in terms of the (generalized) context model developed by Medin and Schaffer (1978) and Nosofsky (1986).

The findings therefore present theoretical and methodological challenges to existing research. Cognitive linguistic treatments of polysemy have traditionally proceeded on the assumption that polysemous words are linguistic categories that can be modelled in a manner akin to the prototype model developed by Rosch and her colleagues (Rosch and Mervis, 1975; Rosch et al., 1976; Rosch, 1978). The findings made in this thesis are difficult to explain under a prototype-based account of word sense representation. What makes application of prototype theory to explain the findings reported here difficult is evidence that categorisation decisions are subject to selective attention effects. This means that a single stimulus can belong to more than one category depending on what characteristic(s) is being used to categorise it. Recall that, unlike the exemplar model, the prototype model neither predicts, not seems to be able to accommodate, selective attention effects and dynamic categories.

Furthermore, close analyses of polysemous words, such as the canonical case studies of *over* by Brugman and Lakoff (2006 [1988]) and Tyler and Evans (2001) offer proposals as to what the senses, and central sense, of polysemous words are *as they have been identified by the authors*. The finding that expert intuitions about what the

senses of a polysemous word are do not always correspond to those of other native speakers, and the subsequent findings that indicated individual differences in word senses, leave the status of these analyses in doubt. They present methodological concerns and urge caution in the continued use of intuition as the means for determining the senses of a polysemous word. The reason for this caution is not only because of the finding that individuals appear to have different senses of polysemous words, but that the theory of their representation, i.e., a theory influenced by the prototype model, does not appear to fit the findings shown here. A prototype theory would allow an intuition-based analysis as long as it acknowledged that the intuitions were those only of the author, and that no commitment as to the reality of those senses in the minds of other speakers was made. However, the observations favour an account of sense representation akin to the exemplar model. In this model, categorisation decisions – i.e., what the sense an exemplar of a given word is – is subject not only to inter-speaker variation, but also *intra-speaker* variation; i.e., the same speaker may classify an exemplar of a polysemous word differently on different occasions. Accordingly, the decisions made by authors of existing analyses of polysemous words may, if asked to do the same again today, reach different conclusions.

5.8 Original contributions to knowledge

The aim of this thesis was to study the nature of word senses as examples of linguistic categories, and to focus in particular on certain aspects of their psychological status. It makes four primary original contributions to knowledge.

The first original contribution to knowledge is the finding is that I failed to reliably agree with naïve and expert participants about how examples of polysemous words should be categorised. This is an original contribution to knowledge about the utility of expert intuitions in the analysis of word senses. It has practical implications in that it discourages finely-grained approaches to word sense distinctions. This may be of use to computational linguists aiming to develop a high quality automated word sense disambiguation algorithm. It problematises the status of existing intuition-based analyses of polysemous words, such as those by Brugman and Lakoff (2006 [1988]), Evans and Tyler (2005), Mahpeykar and Tyler (2011), Masi (2010), and Tyler and Evans (2001). It was speculated that while fine-grained senses identified

by individual speakers may not correspond to each other, coarse grained-senses might.

It was further speculated that our disagreement was due to the possibility that the participants shared a set of senses that were different to mine. This was tested by explicitly assessing whether participants would agree with each other. In this way, I aimed to study whether there are individual differences in word senses. Evidence that participants did not reliably agree with each other over how examples of polysemous words should be sorted indicates that there may indeed be individual differences in word senses. This is an original contribution to knowledge about individual differences in language, and extends the scope of existing research on this topic by taking the issue of individual differences into the domain of semantics. The finding that there appear to be individual differences in word senses entails that, in the first experiment, participants and I disagreed not because my senses did not correspond to a set of senses shared by the participants, but simply because I, like the participants, are individuals with different notions of what constitutes a distinct sense of a given polysemous word.

The thesis also aimed to investigate whether or not word senses are stored in memory. Statistical analyses of how well participants agreed with themselves over how a set of sentences should be sorted demonstrated that participants agree significantly better, and more often, than with other participants. However, the observation that some participants failed to agree with themselves better than with other participants allowed only a tentative conclusion that they are stored in memory, and it was instead suggested that they may have *some form* of representation, but perhaps not a fixed one. This is an original contribution to our knowledge about the storage of word senses. It suggests that polysemous word senses do not have a fixed representation, as suggested by Tyler and Evans (2001), but nor do they have no representation, as implied by the monosemy account in which words have a unitary meaning which is "fleshed out" in context (e.g., Ruhl, 1989; Vandergucht, Willems, and Decuypere, 2007).

The exemplar theory of categorisation can account for both the finding that, by and large, participants agreed with themselves significantly better, and more often, than

with other participants, and the finding that some participants agree with other participants better than they do with themselves. This is accommodated by the fact that the exemplar theory predicts dynamic categories which are constructed in response to the particular demands of a given categorisation task. In exemplar-based models, in a categorisation task one or more categorisation criteria are selectively attended to. Differences among the stimuli in these criteria are more salient, and differences in irrelevant criteria are less salient. If a participant used different criteria to categorise the stimuli at T2 than they did at T1, their categorisation decisions will therefore differ. Given that the instructions given to participants were identical, it is not anticipated that the sorting criteria that participants selectively attend to should change dramatically between T1 and T2. While in general it seems that they did not, this was not always the case. I investigated whether or not the exemplar theory can account for the representation of word senses by testing one of the central predictions of the theory: that categorisation decisions are subject to selective attention. Significant selective attention effects were found. This is an original contribution to knowledge about how exemplar theory can be applied to linguistic categories. It extends existing work on exemplar-based approaches to linguistic categorisation by specifically testing for selective attention effects in semantic categorisation. This suggests that we do not need a special case of the exemplar model to account for word sense categories, contrary to Bybee's (2006) proposal that linguistic categories are so frequent that they have entrenched representations that are not modulated by selective attention.

These two findings, made using evidence gathered from a large sample of naïve participants, suggests that word senses have a non-fixed representation, and may be best understood as *potential* categories of exemplars in a manner compatible with the exemplar model of categorisation. In this way, when we saw evidence of poor intra-participant agreement in Experiment 3, what we were witnessing might have been the dynamic character of linguistic categories. In describing senses as categories, I move away from accounts made by scholars such as Tyler and Evans (2001) and Brugman and Lakoff (2006 [1988]) that polysemous words are categories.

I therefore make four primary original contributions to knowledge and have achieved the aim of the thesis. Collectively, these make original contributions to knowledge about the psychological status of word senses.

5.8.1 Secondary contributions

In the course of making the primary contributions to knowledge described above, I have made two additional, secondary contributions.

First, I have developed an original methodology for studying individual differences in word senses and for studying how they may be represented in the mind. This comprises sentence-sorting tasks in combination with appropriate statistical analyses and network visualisation.

Second, I have offered an original insight into how we can interpret modularity values used to measure network quality. By finding a strong and significant correlation between modularity values and interpretable agreement magnitude values, I have concluded that a modularity value of approximately 0.67 or higher may be considered acceptable.

5.9 Limitations

In this thesis I have made claims about the representation of word senses in memory, arguing that there is evidence that they have a form of storage rather different to that which has been proposed elsewhere. Specifically, I argue that they have neither *no* mental representation, nor a *fixed* representation, but instead that they are *potential categories* of exemplars, and it is these exemplars that are stored in memory. I have reached this conclusion by assessing whether, when asked to categorise sentences exemplifying a particular polysemous word according to the meaning of that word, participants agree with themselves better than with other participants. I predicted that if word senses *do* have some form of non-fleeting mental representation, participants should agree with themselves significantly better, and more often, than with other participants. This was found to be the case.

This is just one approach to studying the representation of word senses, and it is possible that other methodologies may reach different conclusions. These findings

therefore produce tentative conclusions that require further testing using alternative methodologies.

The thesis aims to address only certain aspects of the psychological status of word senses: their storage, what theoretical models can and cannot account for their storage, and individual differences in word senses. There is much more that can be said about word senses, such as how we access word senses in context; what features of context are attended to when reorganizing the multidimensional space – and indeed, why those features are selectively attended to at the expense of others; how children acquire the senses of polysemous words; and if and how senses are related to each other. A large literature tackling these and other aspects of polysemy and word senses exists, and it is intended that this thesis contributes to this rich body of work.

Finally, this thesis studies just four words, all of a closed class nature. At this point, it is therefore difficult to know whether the claims I have made about the psychological status of the senses of these words are generalisable to other words within and beyond the closed class.

5.10 Future research

The length of a doctorate, both temporal and spatial, necessitates that some interesting outcomes requiring further study, and avenues inviting exploration, are not pursued. I outline here some issues that I judge to be particularly worthy of attention in future research.

5.10.1 Accounting for individual differences in word senses

One of the primary findings reported in this thesis is that participants did not reliably agree with each other about how examples of a particular polysemous word should be categorised. This indicates that participants may have different senses of polysemous words. I have not determined why some participants agreed with each other more than with others. In a study of human word sense disambiguation using a method analogous to the closed sort task used in Experiment 1, Murray and Green (2004) found that individuals with different levels of lexical ability, as measured using a component of the Graduate Record Examinations (GRE), disagreed with each other more than with individuals with similar levels of lexical ability. We might

therefore predict that participants with different educational backgrounds will disagree with each other about what the senses of a given word are. However, in Murray and Green's study, all participants were studying for a qualification of bachelor's level or higher. This suggests that even within high educational attainment groups, variation exists. If differences are due to lexical ability, which may be a result of differences in exposure to language, we might predict that individuals in different occupations may disagree with each other more than those who are in similar occupations. For example, I would anticipate that a participant who is employed as a manual labourer may make different sorting decisions than a participant who is employed as a secretary, due to differences in the extent of language exposure across these two occupations. I predict that there would be particularly marked differences in exposure to written texts, which may influence different classification of TEXT USES of *above* and *below*. It is possible that the perceived *meanings* of these uses may differ according to differences in exposure to written English; while it was observed that TEXT USES of *above* appear to have a temporal meaning, and TEXT USES of *below* have a spatial meaning, perhaps individuals with different levels of exposure to written English will reach different conclusions. Differences in language variety may also exist: as was observed in section 2.7.2.2, there were systematic differences in how British English versus American English participants interpreted certain examples of *over*. This suggests that speakers of the same variety may agree with each other more than with speakers of other varieties, at least in the case of some usages.

5.10.2 The use of network visualisations

5.10.2.1 In computational linguistics

In sections 4.3.4, 4.6.3.2, and 4.7.3.4 I briefly discussed the utility of network visualisations in the study of word senses. I noted that they can be studied qualitatively, to analyse community membership and relationships (and absence thereof) between communities. The "quality" of the network can be measured using the modularity value of the network; in the case of this research, this value represents how well the network reflects the categorisation decisions of all participants. I found a strong and significant relationship between the mean inter-participant agreement value for each task and the modularity value of each resultant network. This, and the fact that the communities identified by the algorithm had comprehensible and

seemingly non-random membership indicates that they may be a useful means of studying word senses and their inter-relations. Some networks produced in this research had very low modularity values, indicating that the communities within the network were not useful indicators of word senses. Future research may, however, make fruitful conclusions in the study of both high quality and poor quality networks. Ide and Wilks (2007) have argued that coarse-grained, homograph-like senses may be the most useful level of meaning for successful communication. It might be the case that smaller communities identified by network algorithms can be collapsed to form the coarse-grained senses that Ide and Wilks advocate. A line of research in word sense disambiguation might therefore involve the kind of open-sort tasks used in this study, producing data which is used to create network visualisations. The communities in these networks can then be collapsed into larger "mega-communities", which are then labelled. These labels can be used to tag the member sentences of each mega-communities, which are in turn used as training materials for an automated WSD algorithm. If, when the algorithm is tested using novel material, the algorithm performs acceptably, that would suggest that networks are a useful tool for identifying appropriate word sense distinctions and tags for WSD algorithms. If we understand communities to reflect word senses, the fact that networks generated for some words had high modularity values, and therefore distinct, well-defined communities suggests that senses of some words are, likewise, distinct and well-defined. Equally, networks for some words had low modularity values and produced indistinct, undifferentiated communities, suggesting that senses of those words are also indistinct and poorly defined. Returning to Ide and Wilks' comment about the utility of fine- versus coarse-grained senses in achieving successful communication, and in the face of high agreement values for some words and low agreement values for others, I wonder whether the level of coarseness in sense distinction necessary for successful communication varies across words. For example, we saw in Table 20 that participants in the *under* mixed task at T1 reached a mean agreement score of 0.249, whereas participants in the *over* mixed T1 task reached a mean agreement score of 0.787. If the participants in the *over* mixed T1 task had used a similarly low number of groups, this might explain why agreement was so high. This was not the case; instead, the mean number of groups for this task was 6.8, versus 5.3 for the *under* mixed T1 task. Accordingly, it seems that participants can reach a very high degree of inter-participant agreement even when

they use finely-grained sense distinctions to categorise examples of *over*, but the same is not true of *under*. In terms of Ide and Wilks' argument, it may be that accurate interpretation of a use of *under* depends on accessing a coarsely-grained sense, whereas to correctly interpret a use an *over*, a more finely-grained sense is accessed. Further investigation of this possibility is needed. Moreover, exactly *why* some polysemous words are categorised with better agreement than others promises to be an interesting topic of investigation.

5.10.2.2 For developing theories of lexical representations

While it was not possible to use network visualisations in this particular study of selective attention effects in word sense disambiguation, nor to directly test whether or not word senses are stored in memory, I propose that they may yet prove to be a useful and highly versatile tool for studying lexical representations in cognitive linguistics. In the following section, I set out below three potential uses of network visualisations in this field. I do not claim that these represent the extent of opportunity for their use; instead, these are some potential uses that have emerged following the present research.

5.10.2.2.1 Understanding relationships between the senses of polysemous words

Networks are used in canonical polysemy literature to describe the organisation of word senses; specifically, they are used to specify senses (represented by nodes) and the relationships between them (represented by edges) (e.g., Brugman and Lakoff, 1988; Tyler and Evans, 2001). Their use is motivated by the notion that polysemous words have distinct but *related* senses; networks therefore provide an efficient means of specifying those relations. However, these networks are not uncontroversial; as Sandra and Rice (1995) noted, how well they correspond to mental representations is uncertain. One reason for this uncertainty is their provenance, as networks are typically constructed by the author(s) following their intuitions.

Given their potential for illustrating the connections between what are claimed to be related senses, it seems that networks should indeed play a role in representing relations among word senses. Networks based on sentence-sorting data, such as those presented in Appendix 7 of this thesis, offer a solution to the objections made about the reality of sense relations that scholars have proposed: rather than being intuition-led, they are produced on the basis of data gathered from large samples of

naive participants who are not aware of the purpose of the task. Further, as they are based on categorisation data, and as those categories that participants create are assumed to reflect the mental representations underlying those categories, we can tentatively infer that good quality networks might too correspond with mental representations. In this way, networks produced using data from sentence-sorting tasks not only address Sandra and Rice's valid methodological concerns, but they are also more likely to correspond to the mental representations underlying words and their senses.

Throughout this thesis I have presented a body of evidence indicating that there may be individual differences in the way that people classify sentences featuring a common polysemous word; from this, I concluded that there is evidence of individual differences in word senses. Whether participants reach "acceptable" but not perfect agreement or not, we would expect a certain degree of disagreement over how pairs of sentences should be sorted. It is from such disagreement, represented in the network as inter-community edges, that we might be able to understand which senses are related to others. Let us take the network presented in Figure 43 as an example.[10]

[10] To view the network in full size, please see Appendix 7.

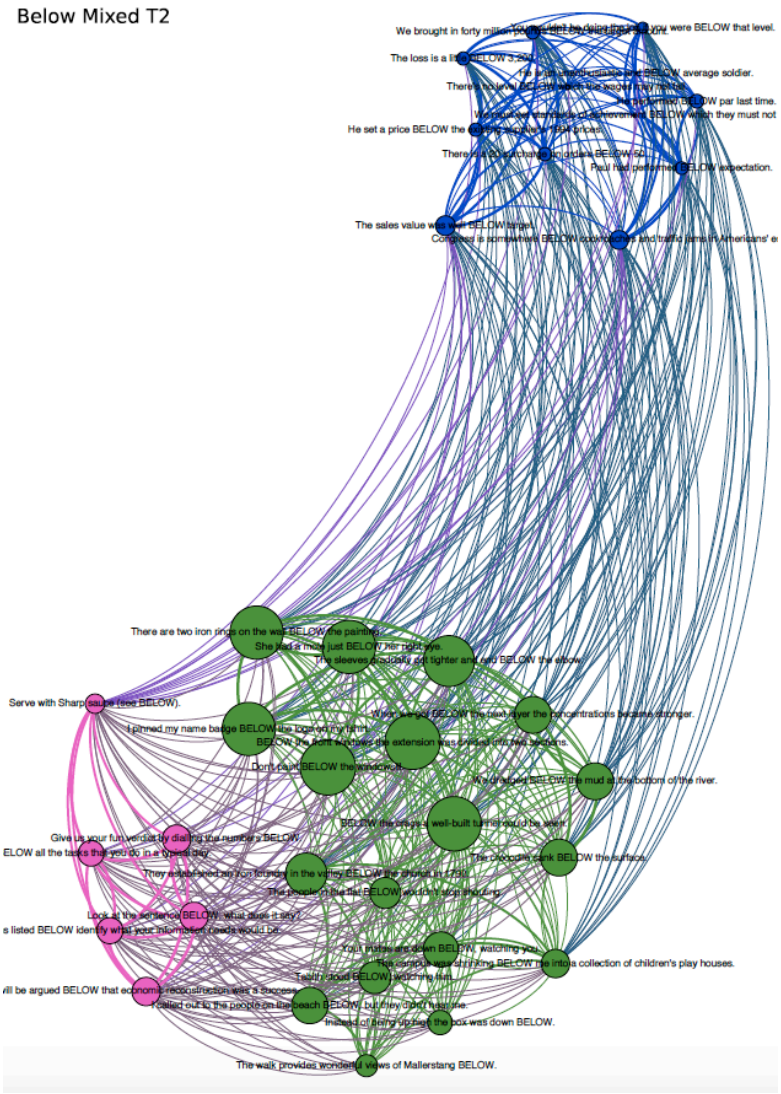


Figure 43 Network visualisation of sentence-sorting task for *Below* mixed sense-type task at T2

This network is taken from the *below* mixed sense-type task at T2. If you attend closely to the three communities, represented in pink, green, and blue, you will note that, while they are distinct communities, there remain connections between individual nodes in pairs of communities. For example; the usage *The campus was shrinking below me into a collection of children's play houses* is a member of the green community, which seems to capture a very generic spatial sense, and is also connected to nodes in the other two communities. These connections represent some participants' decisions that the pair of sentences connected by an inter-community edge belongs in the same group. Their membership of different communities confirms that this was a decision made by the minority, but nonetheless, some still

made that judgment. Such judgments, that pairs of sentences that most participants think exemplify the distinct senses are judged by a minority group to exemplify the same sense, are what might allow us to use networks to identify connections between senses. If a sense were so distinct from other senses that no participant sorted any of its exemplars with exemplars of any other senses, it would manifest as a distinct network. But where the distinction between two senses is less clear, which may be the result of a relationship between those two senses, participants may make different decisions to each other, resulting in communities, reflecting senses, being connected.

As well as showing what pairs of senses are related to each other, networks can also show the overall degree of relatedness of the senses of a polysemous word. Figure 43 shows that, in this case, every community is related to every other community. If we understand communities to reflect senses, we can conclude from this network that the all three senses represented here are related to each other. However, if we consider the network in Figure 44, we see a different outcome. This network, created using data from the *over* non-spatial task at T2, reveals that while each sense appears to be related to at least one other sense, not all senses are related to every other sense.

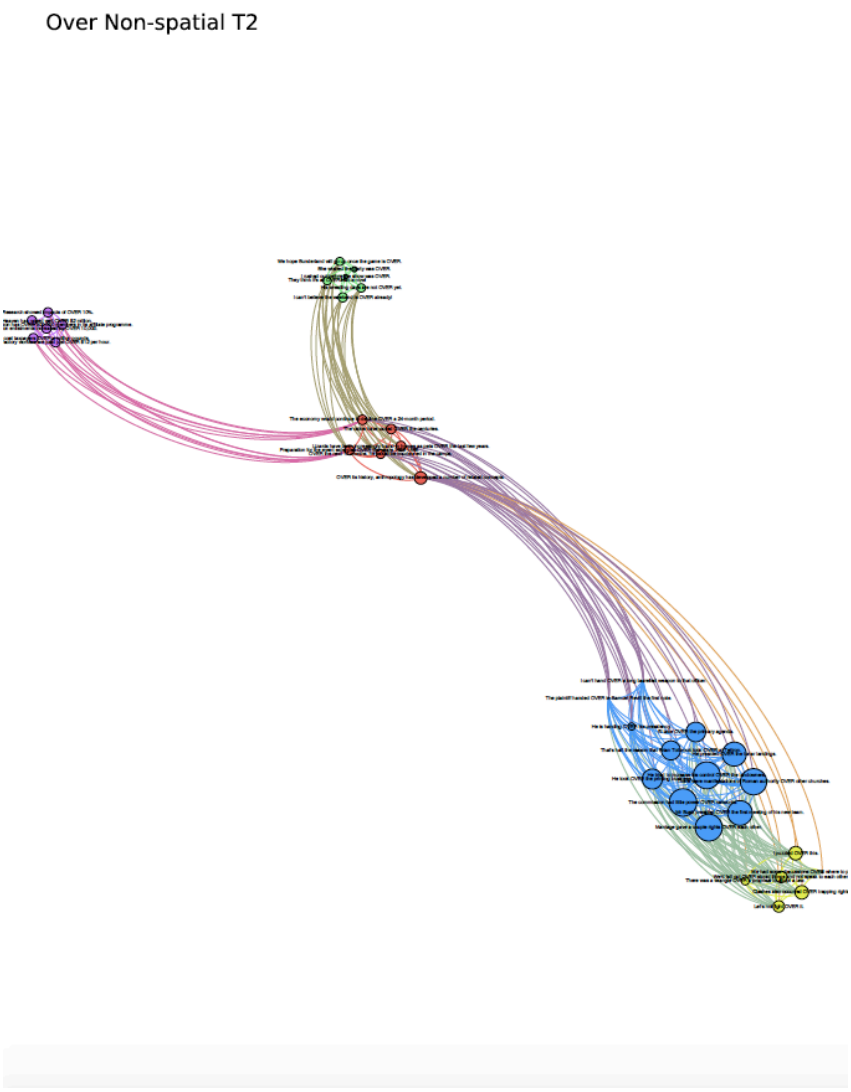


Figure 44 Network visualisation of sentence-sorting data from the *over* non-spatial sense task at T2

These two examples demonstrate the capacity networks based on sentence-sorting data have for providing support for the claim that underpins the theory of polysemy: that word senses are related to each other.

5.10.2.2.2 Distinguishing between homonymous forms and polysemous senses

Just as network visualisations of sentence-sorting data can reveal relationships among senses, they might also reveal which senses are *unrelated* to other senses. As noted above, the senses of polysemous words are related to each other, and in this way word senses differ from meanings of homonyms, which are unrelated. By examining networks to identify communities and connections between communities, there exists, therefore, the opportunity to use network visualisations of data from sentence-sorting tasks to establish not only which senses are related to each other,

and how, but to isolate *meanings*, which are represented in a network visualisation as a community forming a distinct network. Likewise, it may be possible to visualise whether any sets of distinct multi-community networks emerge from sentence-sorting data, which could be taken to represent the senses of unrelated homonyms.

An indication that such an outcome might be found comes from the network produced for the *over* mixed sense-type task at T2, shown in Figure 45.

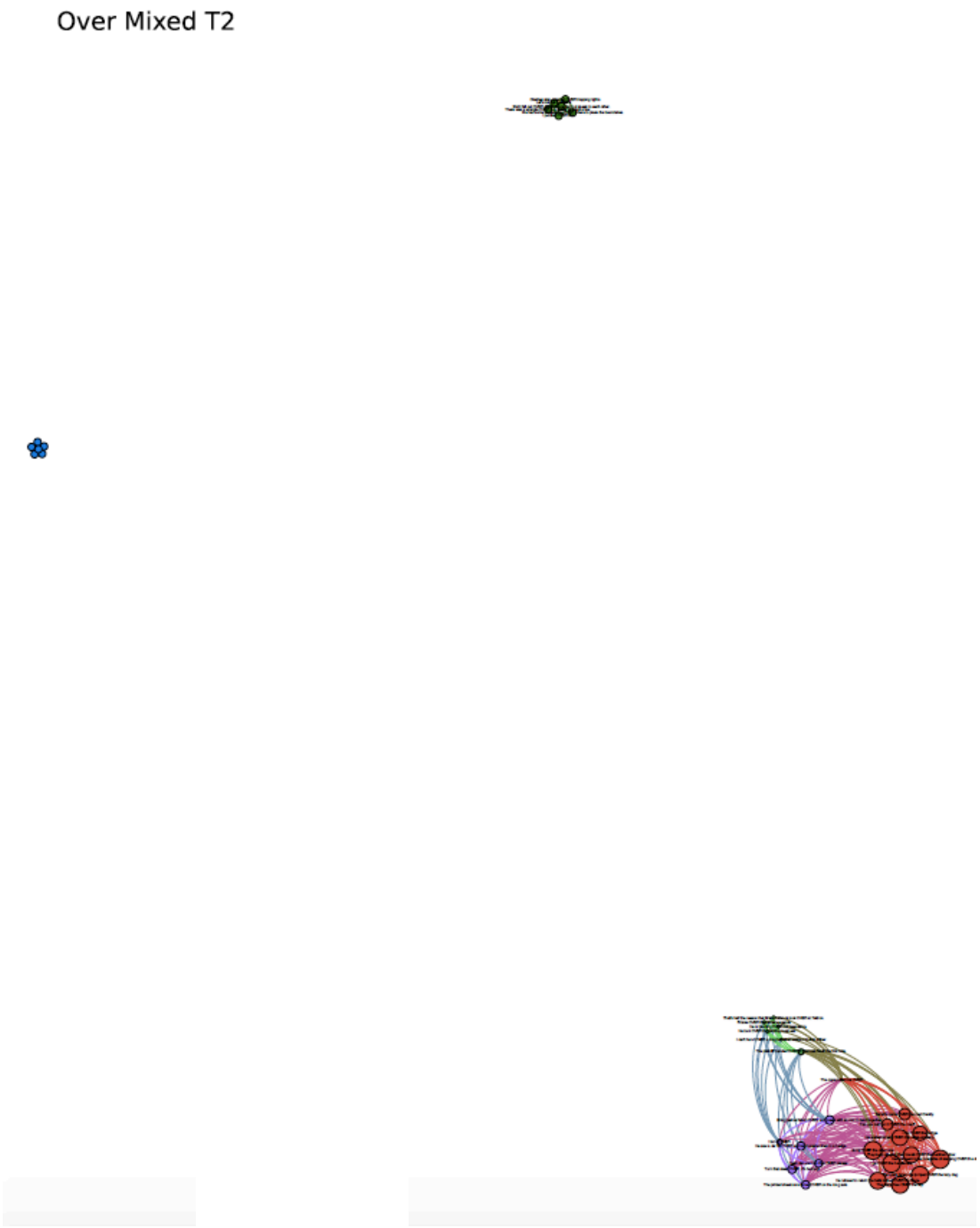


Figure 45 Network visualisation of sentence-sorting data from *over* mixed sense-type task at T2

A single visualisation of the sentence-sorting data produced three networks: a large network featuring three communities, and two additional, single-community networks. On the grounds that polysemous senses and homonymous meanings are broadly distinguished by whether or not they are related to each other, and given the degree of attention that has been paid to the polysemy of *over* over the last thirty years, it is quite surprising to note that two communities, taken to represent senses, are distinct from both each other and the third, larger network. This is an early indication that the single-community networks might not represent senses of the same polyseme as those in the larger network, but might instead reflect meanings of two homonymous forms of *over*. The possibility remains, however, that these two communities *would* be connected (at least by some degree of separation) to the other communities in the visualisation if they were connected to one or more intermediate communities. A very large-scale sorting task, representing as close to a representative sample of the uses of *over* as is conceivable in a sorting task, may shed further light on whether, by some degree, these distinct networks are ever connected to each other, or whether they are unrelated, in which case we may tentatively conclude that the unconnected community represents a homonymous meaning.

5.10.2.2.3 Investigating degrees of near-synonymy

Near synonymy is the phenomenon whereby a single meaning, or two very similar meanings, is expressed using more than one word (Divjak 2010). Divjak observes that while what Cruse (2010) refers to as "cognitive synonyms" are selected according to expressive (e.g., stylistic) demands, near-synonyms are selected according to propositional demands. Take *slice* and *chop*, for example: while both words capture an action carried out using a knife, a sous chef might find herself surprised when instructed to *slice the lettuce* and *chop the bread*: an instruction to *slice the bread* and *chop* the lettuce would feel less semantically anomalous. Understanding words as linguistic categories, if one were to ask participants to sort examples of pairs of polysemous near-synonyms according to their meaning, we could expect the data, when visualised using a network visualisation algorithm, to return at least two networks, with each network capturing examples of just one of the two synonyms. To my knowledge, while sorting tasks have been used in studies of near-synonymy (Divjak and Gries, 2008), network

visualisations of data gathered from sentence-sorting tasks using examples of near-synonyms as stimuli have not been produced, meaning that this prediction remains untested.

Just as network visualisations of sentence sorting data can reveal insights into the relationships – or lack thereof – between the senses of a polysemous word, they may also prove useful in revealing relationships between near-synonyms and their senses. While near-synonymy is a topic of contemporary research in cognitive linguistics (e.g., Divjak & Gries, 2008), to my knowledge, scholars have not yet addressed the possibility that near-synonyms may vary in how synonymous they are, depending on which sense of each synonym is being compared. For example, are *under* and *below* more synonymous in *He felt warm under the blanket* and *The crocodile sank below the surface* than in *We're under real pressure* and *Don't paint below the windowsill*? Working again on the assumption that words and/or their senses are examples of linguistic categories, and that these categories can be observed in networks and communities, we might expect network visualisations of sorting data from a task involving senses of two near-synonyms to produce at least two networks: one for each near-synonym. However, given their similarity in meaning, we might find networks – and indeed communities – that feature *both* near-synonyms. By analysing the networks that emerge from a sorting task of this type, we can understand whether, while similar in meaning, pairs of near synonyms are sufficiently distinct to produce distinct networks. Alternatively, we might observe that this semantic similarity results in communities featuring one synonym that are connected to other communities featuring its synonym pair, or even communities featuring both near-synonyms. Should we observe networks or communities featuring both near-synonyms, by closely analysing what senses are represented by the pairs of connected mono-synonym communities, and/or what sense is represented by dual-synonym communities, we can establish the (varying) degree to which a pair of near-synonyms are synonymous.

5.10.3 Selective attention in linguistic categorisation

One of the principal findings of this thesis is that there appear to be selective attention effects in the particular case of linguistic categorisation that I examined. The narrow scope of the thesis means that I cannot assess whether selective attention

is an effect general to all cases of linguistic categorisation. This will only be determined by work that aims to test for the presence of this effect in other linguistic categories, including other polysemous words.

5.10.4 Word senses as linguistic categories

Finally, the proposal that it is word senses, rather than words themselves, may be linguistic categories contradicts existing theoretical accounts of polysemy. Given that this novel suggestion is made following mere doctoral study, and in light of the fact that I did not explicitly set out to test the hypothesis that polysemous words are not categories, I recommend that this conclusion is tested further.

5.11 Summary

I close this thesis with a summary of its primary original findings, made using original data gathered by three sets of sentence-sorting tasks focusing on four words. First, it does not appear that my sense distinctions align with those of other speakers. This is explained by my second finding, that individuals appear to have different senses of polysemous words. Third, there is evidence that word senses may have some form of mental representation, but not in the fixed form previously advocated (e.g., Tyler and Evans, 2001). Finally, categorisation of exemplars of polysemous words appears to be subject to selective attention effects, allowing the tentative conclusions that senses are represented in a manner consistent with the exemplar model of categorisation, and that we may understand word senses as potential categories of exemplars.

List of references

Albright, A. & Hayes, B., 2003. Rules vs. analogy in English past tenses: a computational/ experimental study. *Cognition*, 90(2), pp.119–161.

Ambridge, B. & Goldberg, A.E., 2008. The island status of clausal complements: Evidence in favor of an information structure explanation. *Cognitive Linguistics*, 19(3), pp.357–389.

Arppe, A. & Järvikivi, J., 2007. Take empiricism seriously! In support of methodological diversity in linguistics. *Corpus Linguistics and Linguistic Theory*, 3(1), pp.99–109.

Artstein, R. & Poesio, M., 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), pp.555–596.

Baetu, I. & Shultz, T.R., 2010. Development of prototype abstraction and exemplar memorization. In *Proceedings of the 32nd annual conference of the cognitive science society*. pp. 814–819.

Baker, C.F., 1999. *Seeing clearly: Frame Semantic , Psycholinguistic, and Cross-linguistic Approaches to the Semantics of the English Verb See*. University of California, Berkeley.

Barlow, M. & Kemmer, S., 2000. *Usage Based Models of Language*, Cambridge: Cambridge University Press.

Bates, E., Dale, P.S. & Thal, D., 1995. Individual differences and their implications for theories of language development. In P. Fletcher & B. MacWhinney, eds. *Handbook of Child Language*. Oxford: Basil Blackwell.

Beitel, D.A., Gibbs, R.W.J. & Sanders, P., 1997. The embodied approach to the polysemy of the spatial preposition on. In H. Cuyckens & B. Zawada, eds. *Polysemy in Cognitive Linguistics: Selected Papers from the International Cognitive Linguistics Conference, Amsterdam, 1997*. Amsterdam: John Benjamins, pp. 241–260.

Berez, A.L. & Gries, S.T., 2008. In defense of corpus-based methods: A behavioral profile analysis of polysemous get in English. In S. Moran, D. S. Tanner, & M. Scanlon, eds. *Proceedings of the 24th Northwest Linguistics Conference*. Seattle, WA: University of Washington, pp. 157–166.

Berlin, B., Breedlove, D.E. & Raven, P.H., 1974. *Principles of Tzeltal Plant*

Classification, New York: Academic Press.

Bhala, R.V.V. & Abirami, S., 2014. Trends in word sense disambiguation. *Artificial Intelligence Review*, 42(2), pp.159–171.

Bhardwaj, V. et al., 2010. Anveshan: a framework for analysis of multiple annotators' labeling behavior. *Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV)*, pp.47–55.

Blondel, V.D., Guillaume, J. & Lefebvre, E., 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10, p.P10008.

Bradac, J.J. et al., 1980. On the neglected side of linguistic science: multivariate studies of sentence judgment. *Linguistics*, 18(11-12), pp.967–995.

Brugman, C., 1981. *The story of over*. University of California, Berkeley.

Brugman, C. & Lakoff, G., 1988. Cognitive topology and lexical networks. In S. L. Small, G. W. Cottrell, & M. K. Tanenhaus, eds. *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence*. San Mateo, CA: Morgan Kaufmann, pp. 477–508.

Brugman, C. & Lakoff, G., 2006. Cognitive topology and lexical networks. In D. Geeraerts, ed. *Cognitive Linguistics: Basic Readings*. Berlin: Mouton de Gruyter, pp. 109–139.

Bybee, J., 2006. From usage to grammar: The mind's response to repetition. *Language*, 82(4), pp.711–733.

Bybee, J., 2002. Phonological evidence for exemplar storage of multiword utterances. *Studies in Second Language Acquisition*, 24(2), pp.215–221.

Carletta, J., 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2), pp.249–254.

Chandler, S., 2015. The analogical modeling of linguistic categories. *Language and Cognition*, (October 2015), pp.1–36.

Chandler, S., 2010. The English past tense: Analogy redux. *Cognitive Linguistics*, 21(2010), pp.371–417.

Clark, H., 1983. Making sense of nonce sense. In G. Flores d'Arcais & R. Jarvella, eds. *The Process of Understanding Language*. New York, NY: Wiley, pp. 297–331.

Coleman, L. & Kay, P., 1981. Prototype semantics: The English word lie. *Language*, 57(1), pp.26–44.

Croft, W. & Cruse, D.A., 2004. *Cognitive Linguistics*, Cambridge: Cambridge University Press.

Cuyckens, H., Sandra, D. & Rice, S., 1997. Towards an empirical lexical semantics. In B. Smieja & M. Tasch, eds. *Human Contact Though Language and Linguistics*. Berlin: Peter Lang, pp. 35–54.

Cuyckens, H. & Zawada, B., 2001. *Prepositions in Cognitive Linguistics: selected papers from the International Cognitive Linguistics Conference, Amsterdam, 1997*, Amsterdam: John Benjamins.

Dąbrowska, E., 2012. Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism*, 2(2), pp.219–253.

Dąbrowska, E., 2010. Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review*, 27(1), pp.1–23.

Dąbrowska, E., 2008. Questions with long-distance dependencies: A usage-based perspective. *Cognitive Linguistics*, 19(3), pp.391–425.

Dąbrowska, E. & Street, J., 2006. Individual differences in language attainment: Comprehension of passive sentences by native and non-native English speakers. *Language Sciences*, 28(6), pp.604–615.

Daelemans, W. et al., 2002. *TimBL: Tilburg Memory-Based Learner, version 4.3 reference guide*, Tilburg: ILK.

Divjak, D., 2003. On trying in Russian: a tentative network model for near(er) synonyms. "Belgian Contributions to the 13th International Congress of Slavicists, Ljubljana, 15-21 August 2003". *Slavica Gandensia*, 30, pp.25–58.

Divjak, D., 2010. *Structuring the Lexicon: A Clustered Model for Near-Synonymy*, Berlin/New York: Walter de Gruyter.

Divjak, D. & Arppe, A., 2013. Extracting prototypes from exemplars What can corpus data tell us about concept representation? *Cognitive Linguistics*, 24(2), pp.221–274.

Divjak, D., Dąbrowska, E. & Arppe, A., 2016. Machine Meets Man: Evaluating the psychological reality of corpus-based probabalistic models. *Cognitive Linguistics*, 27(1), pp.1–33.

Divjak, D. & Gries, S.T., 2008. Clusters in the mind? Converging evidence from near-synonymy in Russian. *The Mental Lexicon*, 3(2), pp.188–213.

Duffy, S.E., 2015. *The metaphoric representation of time: a cognitive linguistic*

perspective. Northumbria University.

Duffy, S.E. & Feist, M.I., 2014. Individual differences in the interpretation of ambiguous statements about time. *Cognitive Linguistics*, 25(1), pp.29–54.

Duffy, S.E., Feist, M.I. & McCarthy, S., 2014. Moving through time: The role of personality in three real life contexts. *Cognitive Science*, 38(8), pp.1662–1674.

Durkin, K. & Manning, J., 1989. Polysemy and the subjective lexicon: semantic relatedness and the salience of intraword senses. *Journal of Psycholinguistic Research*, 18(6), pp.577–612.

Ellis, N.C., 2006. Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, 27(2), pp.164–194.

Erk, K., McCarthy, D. & Gaylord, N., 2009. Investigations on word senses and word usages. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 1, pp.10–18.

Evans, V. & Tyler, A., 2005. Applying Cognitive Linguistics to pedagogical grammar: the English prepositions of verticality. *Revista Brasileira de Linguistica Aplicada*, 5(2), pp.11–42.

Evans, V. & Tyler, A., 2004. Spatial experience, lexical structure and motivation: The case of in. In G. Radden & K.-U. Panther, eds. *Studies in Linguistic Motivation*. Berlin/New York: Mouton de Gruyter, pp. 157–192.

Fernald, A. & Marchman, V.A., 2012. Individual differences in lexical processing at 18 months predict vocabulary growth in typically-developing and late-talking toddlers. *Child Development*, 83(1), pp.203–222.

Ferreira, F., Bailey, K.G.D. & Ferraro, V., 2002. Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11(1), pp.11–15.

Foraker, S. & Murphy, G.L., 2012. Polysemy in sentence comprehension: Effects of meaning sominance. *Journal of Memory and Language*, 67(4), pp.407–425.

Francis, A.L. & Nusbaum, H.C., 2002. Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, 28(2), pp.349–366.

Gahl, S. & Yu, A.C.L., 2006. Introduction to the special issue on exemplar-based models in linguistics. *The Linguistic Review*, 23(3), pp.213–216.

Gibbs, R.W.J., 2006. Introspection and Cognitive Linguistics: Should we trust our own intuitions. In J. Ruiz de Mendoza Ibáñez, ed. *Annual Review of Cognitive Linguistics: Volume 4*. Amsterdam: John Benjamins, pp. 135–151.

Gibbs, R.W.J. & Matlock, T., 2001. Psycholinguistic perspectives on polysemy. In H. Cuyckens & B. Zawada, eds. *Polsemy in Cognitive Linguistics*. Amsterdam: John Benjamins, pp. 213–240.

Gibson, E. & Fedorenko, E., 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, 28(1-2), pp.88–124.

Gilquin, G. & McMichael, A., 2008. Measures of prototypicality: Convergence or divergence. The case of through. In *Third International Conference of the German Cognitive Linguistics Association, University of Leipzig*.

Good, B.H., Montjoye, Y. De & Clauset, A., 2010. The performance of modularity maximization in practical contexts. *Physical Review E*, 81, p.046106.

Gordon, P.C. & Hendrick, R., 1997. Intuitive knowledge of linguistic co-reference. *Cognition*, 62(3), pp.325–70.

Gries, S.T., 2010. Behavioral profiles: A fine-grained and quantitative approach in corpus-based lexical semantics. *The Mental Lexicon*, 5, pp.323–346.

Gries, S.T., 2006. Corpus-based methods and cognitive semantics: The many senses of to run. In *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax And Lexis*. pp. 57–99.

Gries, S.T., 2015. Polysemy. In E. Dąbrowska & D. Divjak, eds. *Handbook of Cognitive Linguistics*. Berlin: De Gruyter Mouton, pp. 472–490.

Gries, S.T. & Divjak, D., 2009. Behavioral profiles: a corpus-based approach to cognitive semantic analysis. In V. Evans & S. Pourcel, eds. *New Directions in Cognitive Linguistics*. Amsterdam: John Benjamins, pp. 57–76.

Gumtree, 2017. Bookcase. *Gumtree*. Available at: https://www.gumtree.com/p/other-dining-living-furniture/bookcase/1252964434 [Accessed July 22, 2017].

Hanks, P., 2000. Do word meanings exist? *Computers and the Humanities*, 34, pp.205–215.

Hong, J. & Baker, C.F., 2011. How good is the crowd at "real" WSD? In *Proceedings of the Fith Law Workshop (LAW V)*. Portland, Oregon: Association for Computational Linguistics, pp. 30–37.

Ibarretxe-Antuñano, I., 2004. Polysemy in Basque locational cases. *Belgian Journal*

of Linguistics, (18), pp.271–298.

Ibbotson, P. et al., 2012. Semantics of the transitive construction: Prototype effects and developmental comparisons. *Cognitive Science*, 36(7), pp.1268–1288.

Ibbotson, P. & Tomasello, M., 2009. Prototype constructions in early language acquisition. *Language and Cognition*, 1(1), pp.59–85.

Ide, N. & Wilks, Y., 2007. Making Sense About Sense. In E. Agirre & P. Edmonds, eds. *Word Sense Disambiguation Algorithms and Applications*. Dordrecht: Springer, pp. 47–73.

James, L.J., 1962. *Effects of repeated stimulation on cognitive aspects of behavior: Some experiments on the phenomenon of semantic satiation*. McGill University.

Kalyan, S., 2012. Similarity in linguistic categorization: The importance of necessary properties. *Cognitive Linguistics*, 23(3), pp.539–554.

Kauffman, J. et al., 2014. DyCoNet: A Gephi plugin for community detection in dynamic complex networks. *PLOS One*, 9(7).

Kilgarriff, A., 1998. Gold standard datasets for evaluating word sense disambiguation programs. *Computer Speech & Language*, 12(3), pp.453–472.

Kilgarriff, A., 1997. I don't believe in word senses. *Computers and the Humanities*, 31, pp.91–113.

Kilgarriff, A., 2007. Word senses. In E. Agirre & P. Edmons, eds. *Word Sense Disambiguation: Algorithms and Applications*. Springer, pp. 29–46.

Kishner, J.M. & Gibbs, R.W.J., 1996. How "just" gets its meanings: Polysemy and context in psychological semantics. *Language and Speech*, 39(1), pp.19–36.

Klein, D.E. & Murphy, G.L., 2002. Paper has been my ruin: conceptual relations of polysemous senses. *Journal of Memory and Language*, 47(4), pp.548–570.

Kövecses, Z., 2002. *Metaphor: A Practical Introduction*, Oxford: Oxford University Press.

Labov, W., 1978. Denotational structure. In D. Farkas, W. M. Jacobsen, & K. W. Todrys, eds. *Parasession on the Lexicon*. Chicago: Chicago Linguistic Society, pp. 220–260.

Labov, W., 1972. *Sociolinguistic Patterns*, Philadelphia: University of Pennsylvania Press.

Lakoff, G., 1990. The invariance hypothesis: is abstract reason based on image-schemas? *Cognitive Linguistics*, 1(1), pp.39–74.

Lakoff, G. & Johnson, M., 1980. *Metaphors We Live By*, Chicago: Chicago

University Press.

Landis, J.R. & Koch, G.G., 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1), pp.159–174.

Langacker, R., 1986. An Introduction to Cognitive Grammar. *Cognitive Sci*, 10(1), pp.1–40.

Langacker, R.W., 2001. Discourse in Cognitive Grammar. *Cognitive linguistics*, 12.2(2001), pp.143–188.

Leonard, L.B., 1980. Individual differences in early child phonology. *Applied Psycholinguistics*, 1, pp.7–30.

Levallois, C., 2013. Modularity score. *Gephi Forums*. Available at: http://gephi.forumatic.com/viewtopic.php?f=32&t=3090 [Accessed May 31, 2016].

Lively, S.E., Logan, J.S. & Pisoni, D.B., 1993. Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *J Acoust Soc Am.*, 94(3 part 1), pp.1242–1255.

LoveAntiques, 2017. Victorian Mahogany Bookcase c.1880. *loveantiques.com*. Available at: https://www.loveantiques.com/antique-bookcases/open-bookcase/mahogany/victorian-mahogany-bookcase-c1880-66656 [Accessed July 22, 2017].

Lupyan, G. & Mirman, D., 2013. Linking language and categorization: Evidence from aphasia. *Cortex*, 49(5), pp.1187–1194.

MacLaury, R.E., 1991. Prototypes revisited. *Annual Review of Anthropology*, 20, pp.55–74.

MacLaury, R.E., 1989. Zapotec body-part locatives: Prototypes and metaphoric extensions. *International Journal of American Linguistics*, 55(2), pp.119–154.

MacWhinney, B., 2000. *The CHILDES Project: Tools For Analyzing Talk*, Mahwah, NJ: Lawrence Erlbaum Associates.

Mahpeykar, N. & Tyler, A., 2015. A principled Cognitive Linguistics account of English phrasal verbs with up and out. *Language and Cognition*, 7(1), pp.1–35.

Mahpeykar, N. & Tyler, A., 2011. The semantics of Farsi bi: Applying the Principled Polysemy model. In M. Egenhofer et al., eds. *Spatial Information Theory*. Berlin/Heidelberg: Springer, pp. 413–433.

Masi, S., 2010. English vs. Italian spatial particles of verticality: Over vs. sopra.

Textus, (3), pp.673–696.

McGlone, M.S. & Harding, J.L., 1998. Back (or forward?) to the future: The role of perspective in temporal language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(5), pp.1211–1223.

Medin, D.L. & Schaffer, M.M., 1978. A context theory of classification learning. *Psychological Review*, (85), pp.207–238.

Miller, G., 1971. Empirical methods in the study of semantics. In D. D. Steinberg & L. A. Jakobovits, eds. *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology*. Cambridge: Cambridge University Press, pp. 569–585.

Miller, G., 1962. *Psychology: The Science of Mental Life*, New York: Harper & Row.

Morey, L.C. & Agresti, A., 1984. The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement. *Educational and Psychological Measurement*, 44, pp.33–37.

Mundy, P. & Gomes, A., 1998. Individual differences in joint attention skill development in the second year. *Infant Behavior and Development*, 21(3), pp.469–482.

Murphy, G.L., 2007. Parsimony and the psychological representation of polysemous words. In M. Rakova, G. Petho, & C. Rakosi, eds. *The Cognitive Basis of Polysemy*. Frankfurt am Main: Peter Lang, pp. 47–70.

Murphy, G.L., 2004. *The Big Book of Concepts*, Cambridge, MA: MIT Press.

Murray, G.C. & Green, R., 2004. Lexical knowledge and human disagreement on a WSD task. *Computer Speech & Language*, 18(3), pp.209–222.

Myachykov, A., Thompson, D., et al., 2011. Visual Attention and Structural Choice in Sentence Production Across Languages. *Linguistics and Language Compass*, 5(2), pp.95–107.

Myachykov, A., Garrod, S. & Scheepers, C., 2011. Perceptual priming of structural choice during English and Finnish sentence production. In R. K. Mishra & N. Srinivasan, eds. *Language and Cognition: State of the Art*. Munich: Lincom Europa Press, pp. 53–71.

Myachykov, A., Tomlin, R.S. & Posner, M.I., 2005. Attention and empirical studies of grammar. *The Linguistic Review*, 22, pp.347–364.

Neuendorf, K.A., 2002. *The Content Analysis Guidebook*, Thousand Oaks,

California: Sage Publications Inc.

Newman, M.E.J., 2012. Communities, modules and large-scale structure in networks. *Nature Physics*, 8(1), pp.25–31.

Nosofsky, R.M., 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115(1), pp.39–57.

Ostermann, C., 2014. Particles in cognitive dictionary entries: The case of over. In *UK-CLC 5: 5th UK Cognitive Linguistics Conference*. Lancaster.

Passonneau, R.J., Bhardwaj, V., et al., 2012. Multiplicity and word sense: Evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46, pp.219–252.

Passonneau, R.J., Baker, C., et al., 2012. The MASC word sense sentence corpus. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, pp.3025–3030.

Passonneau, R.J. et al., 2010. Word sense annotation of polysemous words by multiple annotators. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*. pp. 3244–3249.

Passonneau, R.J., Salleb-Aouissi, A. & Ide, N., 2009. Making sense of word sense variation. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*. Association for Computational Linguistics, pp. 2–9.

Pierrehumbert, J.B., 2000. Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. J. Hopper, eds. *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins, pp. 137–158.

Pine, J.M. & Lieven, E.V.M., 1993. Reanalysing rote-learned phrases: individual differences in the transition to multi-word speech. *Journal of Child Language*, 20(03), pp.551–571.

Revelle, W., 2015. Package "psych." Available at: https://cran.r-project.org/web/packages/psych/psych.pdf [Accessed August 25, 2015].

Rice, S., 1993. Far afield in lexical fields: The English prepositions. In M. Bernstein, ed. *ESCOL '92 Proceedings*. Ithaca: Cornell University Press, pp. 206–217.

Rice, S., 2003. Growth of a lexical network: Nine English prepositions in acquisition. In H. Cuyckens, R. Dirven, & J. R. Taylor, eds. *Cognitive Approaches to Linguistic Semantics*. Berlin: Mouton de Gruyter, pp. 243–280.

Rice, S., 1996. Prepositional prototypes. In M. Pütz & R. Dirven, eds. *The Construal*

of Space in Language and Thought. Berlin/New York: Mouton de Gruyter, pp. 135–165.

Rice, S., Sandra, D. & Vanrespaille, M., 1999. Prepositional semantics and the fragile link between space and time. In M. K. Hiraga, C. Sinha, & S. Wilcox, eds. *Cultural, Psychological and Typological Issues in Cognitive Linguistics*. Amsterdam: John Benjamins, pp. 107–127.

Robinson, P., 1995. Attention, memory, and the "noticing" hypothesis. *Language Learning*, 45(2), pp.283–331.

Rosch, E., Mervis, C.B., et al., 1976. Basic objects in natural categories. *Cognitive Psychology*, 8, pp.382–439.

Rosch, E., 1973. On the internal structure of perceptual and semantic categories. In T. E. Moore, ed. *Cognitive Development and the Acquisition of Language*. New York, NY: Academic Press, pp. 111–144.

Rosch, E., 1978. Principles of categorization. In E. Rosch & B. Lloyd, eds. *Cognition and Categorization*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rosch, E. & Mervis, C.B., 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4), pp.573–605.

Rosch, E., Simpson, C. & Miller, R.S., 1976. Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, 2(4), pp.491–502.

Ross, J.R., 1979. Where's English? In C. J. Fillmore, D. Kempler, & W. S.-Y. Wang, eds. *Individual Differences in Language Ability and Language Behaviour*. New York: Academic Press, pp. 127–163.

Ruhl, C., 1989. *On Monosemy: A Study in Linguistic Semantics*, Albany, New York: State University of New York Press.

Sandra, D. & Rice, S., 1995. Network analyses of prepositional meaning: Mirroring whose mind - the linguist's or the language user's? *Cognitive Linguistics*, 6(1), pp.89–130.

Schütze, C.T., 1996. *The Empirical Base of Linguistics: Grammaticality judgments and linguistic methodology*, Chicago: University of Chicago Press.

Schwarz-Friesel, M., 2012. On the status of external evidence in the theories of cognitive linguistics: Compatibility problems or signs of stagnation in the field? Or: Why do some linguists behave like Fodor's input systems? *Language Sciences*, 34(6), pp.656–664.

Seddon, J.M. et al., 1990. Evaluation of an iris color classification system. *Investigative Opthalmology and Visual Science*, 31(8), pp.1592–1598.

Shore, C.M., 1995. *Individual Differences in Language Development*, Thousand Oaks, California: Sage Publications Inc.

Skousen, R., 1989. *Analogical Modeling of Language*, Dordrecht: Kluwer Academic.

Skousen, R., 1992. *Analogy and Structure*, Dordrecht: Kluwer Academic.

Sloman, S.A. & Rips, L.J., 1998. Similarity as an explanatory construct. *Cognition*, 65, pp.87–101.

Smith, E.E. & Sloman, S. a, 1994. Similarity- versus rule-based categorization. *Memory & Cognition*, 22(4), pp.377–86.

Snow, R. et al., 2008. Cheap and fast - but is it good? Evaluation non-expert annotiations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 254–263.

Snyder, W., 2000. An experimental investigation of syntactic satiation effects. *Linguistic Inquiry*, 31(3), pp.575–582.

Spencer, N.J., 1973. Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of Psycholinguistic Research*, 2(2), pp.83–98.

Sprouse, J. & Almeida, D., 2012. Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics*, 48(3), pp.609–652.

Sprouse, J., Schütze, C.T. & Almeida, D., 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua*, 134, pp.219–248.

Srinivasan, M. & Rabagliati, H., 2015. How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. *Lingua*, 157, pp.124–152.

Stefanowitsch, A., 2011. Cognitive linguistics meets the corpus. In M. Brdar, S. T. Gries, & M. Z. Fuchs, eds. *Converging and Diverging Tendencies in Cognitive Linguistics*. Amsterdam: John Benjamins, pp. 257–290.

Stevens, C., Sanders, L. & Neville, H., 2006. Neurophysiological evidence for selective auditory attention deficits in children with specific language impairment. *Brain Research*, 1111(1), pp.143–152.

Street, J.A. & Dąbrowska, E., 2010. More individual differences in language

attainment: How much do adult native speakers of English know about passives and quantifiers? *Lingua*, 120(8), pp.2080–2094.

Sun, S., 2011. Meta-analysis of Cohen's kappa. *Health Services and Outcomes Research Methodology*, 11(3), pp.145–163.

Talmy, L., 2007. Foreword. In M. Gonzalez-Marquez, I. Mittelberg, & M. J. Spivey, eds. *Methods in Cognitive Linguistics*. Amsterdam: John Benjamins, pp. xi–xxi.

Taylor, J.R., 2003. *Linguistic Categorization* Third., Oxford: Oxford University Press.

Tomasello, M. & Farrar, M.J., 1986. Joint attention and early language. *Child Development*, 57(6), pp.1454–1463.

Tomasello, M. & Todd, J., 1983. Joint attention and lexical acquisition style. *First Language*, 4, pp.197–211.

Tuggy, D., 1999. Linguistic evidence for polysemy in the mind: a response to William Crift and Dominiek Sandra. *Cognitive Linguistics*, 10(4), pp.343–368.

Tyler, A. & Evans, V., 2001. Reconsidering prepositional polysemy networks: The case of over. *Language*, 77(4), pp.724–765.

Vandergucht, F., Willems, K. & Decuypere, L., 2007. The iconicity of embodied meaning. Polysemy of spatial prepositions in the cognitive framework. *Language Sciences*, 29(6), pp.733–754.

Vanpaemel, W. & Storms, G., 2008. In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review*, 15(4), pp.732–749.

Vasilyeva, M., Waterfall, H. & Huttenlocher, J., 2008. Emergence of syntax: commonalities and differences across children. *Developmental Science*, 11(1), pp.84–97.

Verhagen, A., 2015. Grammar and cooperative communication. In E. Dabrowska & D. Divjak, eds. *Handbook of Cognitive Linguistics*. Berlin/Boston: Walter de Gruyter, pp. 232–252.

Véronis, J., 1998. A study of polysemy judgements and inter-annotator agreement. In *Programme and advanced papers of the Senseval workshop*.

Willems, K., 2012. Intuition, introspection and observation in linguistic inquiry. *Language Sciences*, 34(6), pp.665–681.

Wilson, B. et al., 2013. Auditory artificial grammar learning in Macaque and Marmoset monkeys. *The Journal of Neuroscience*, 33(48), pp.18825–18835.

Wittgenstein, L., 1958. *Philosophical Investigations*, New York: Macmillan.

Appendix 1: Experiment 1 task instructions

It may not be immediately obvious, but many words we use have more than one meaning. I am interested in finding out the different meanings of the word [target word]. This word is typically understood as having a spatial meaning. A similar word like IN also has a spatial meaning, like this:

(1) The ball is IN the bowl.

But it also has other meanings, some of which might be spatial, like (2), or metaphorical, like (3):

(2) She's sitting in the dark

(3) Mary is IN love with John

The purpose of these examples is to show you that a word as simple as IN can have lots of different meanings; some are obviously related, some take some thinking about before you can see a connection, while some might seem completely unrelated to others.

Instructions

Your task is to **organise a list of sentences into up to six groups.**

On the next screen you'll see the sentences (on the left) and a set of six groups (on the right), each labelled with an example of [target word] being used in real life. Grouping them is easy:

1. Read through all of the sentences and group labels, and think about what [target word] means in each.
2. Drag one sentence at a time into one of the groups in the right, **so that the meaning of [target word] in that sentence matches the meaning of [target word] in the group label.**
3. You are free to use as many of the given groups as you wish.

If you change your mind about where a sentence should go, or if you drop a sentence into the wrong group, you can drag it and drop it into another group.

When you've finished, click 'Finished' in the top-right corner.

Appendix 2: Popular placement matrices

Over

Popular Placements Matrix

	ABOUT	A-B MOVEMENT (NO ARC)	ARC	TRANSFER	COMPLETION	FLIP
There was a wrangle OVER a prop...	100%					
I puzzled OVER this.	100%					
Let's not fight OVER it.	99%				1%	
We had some discussions OVER w...	99%			1%		
Clashes also occurred OVER trapp...	98%	1%		1%		
We'd fall out OVER stupid things a...	96%		1%	1%	1%	
Can you just run it OVER the road?		91%	8%			1%
He walked slowly OVER the zebra-...		91%	9%			
I ran OVER the bridge.		91%	9%			
Sarah's come OVER the road Dadd...		90%	9%	1%		
The plane flew OVER the city.		53%	48%			
The cops pulled me OVER.	3%	43%	13%	28%	1%	14%
I go OVER the handlebars.		6%	94%			
He refused to return the balls kicke...		8%	91%			1%
Jump OVER the other one.		10%	90%			
The quick brown fox jumped OVER...		13%	88%			
They keep slinging their towels OV...		10%	86%			4%
Herons seem to be incapable of st...	1%	14%	84%	1%		
He took OVER the printing busines...				100%		
That's half the reason that Brian To...				100%		
I'll take OVER the primary agenda.	1%			99%		
He is handing OVER his presidency.	1%	1%		94%		4%
The plaintiff handed OVER to Sam...	1%	5%	3%	75%		16%
I can't hand OVER a long barrelled...	1%	5%	8%	71%		15%
She wished the party was OVER.					100%	
I can't believe the weekend is OVE...					100%	
His wrestling days are not yet OVE...					100%	
We hope Sunderland will go up on...					100%	
I rushed out before the show was ...		1%			99%	
They think it's all OVER...it is now!	4%			1%	95%	
I turn it OVER.		1%				99%
The printed sheets are turned OVE...		1%	1%			98%
Yeah can you turn that OVER pleas...				6%		94%
Turn that steak OVER, it's burning!		1%	3%		4%	93%
He saw a car flip OVER and land u...		4%	13%			84%
Bring pelmet fabric OVER, and pre...	1%	16%	11%	15%	1%	55%

Under

Popular Placement	HORIZONTAL RELATIONSHIP	ACCORDING TO	SUBJECT TO	VERTICAL RELATIONSHIP NO CONTACT	VERTICAL RELATIONSHIP WITH CONTACT	UNDER THE CONTROL / AUTHORITY OF
I'm wearing a vest UNDER this shirt.	98%				2%	
Should I wear a jumper UNDER thi...	97%				3%	
I'm a bit hot UNDER the collar.	62%		24%	6%	6%	2%
They found more wallpaper UNDE...	51%			3%	46%	
You can emigrate to Britain UNDER...		97%	2%			2%
The fridge can be exchanged UND...		97%		2%		2%
It will be contested UNDER the Co...		95%	2%			3%
He will be committed UNDER the m...		95%	2%			3%
They will file an explanation UNDE...		94%	5%			2%
UNDER the new regulations, the st...		94%				6%
They deny they were UNDER any ...		40%	37%			24%
This is something which is very mu...		2%	94%			5%
Remember their hospital is UNDER...		3%	89%			8%
Your application is now UNDER rev...		11%	84%			5%
The question of intercommunion is...		10%	83%			8%
We're UNDER real pressure at the ...		2%	75%		2%	22%
They got UNDER cover of the walls...	13%		43%	24%	14%	6%
They lay UNDER the stars.	2%	2%	2%	87%	8%	
The Troll seldom came out from U...	2%			75%	24%	
Cover dark circles UNDER your ey...	16%			71%	13%	
Rinse the dish UNDER running wat...	3%		5%	65%	27%	
She kicked him UNDER the table.	3%		2%	62%	33%	
I'm hiding UNDER your bed.	3%			56%	41%	
I'm frying the bread UNDER there.	6%		2%	52%	40%	
They put material UNDER the carp...	6%	2%		2%	90%	
The growing roots UNDER the pat...	3%			10%	87%	
They looked where others wouldn'...	3%		2%	16%	79%	
They'd been hiding people UNDER...	8%			25%	67%	
I'm sleeping UNDER my cover.	19%			17%	63%	
He felt warm UNDER the blanket.	32%			13%	56%	
250 Garrimperos work UNDER him.						100%
She's got a whole team working U...			3%			97%
I'm working UNDER the direction o...		2%	3%			95%
He served in 102 Battalion UNDER...		3%	3%			94%
The Act was introduced UNDER th...		17%				83%
We'll be taxed UNDER the Conserv...		22%	3%			75%

Above

Popular Placements Matrix

	Better than	More than	Vertical relationship	Text uses	Vantage	Hierarchy
Are they good, ABOVE average, or...	92%	5%				2%
It was either ABOVE average or be...	91%	9%				
The Renault 5 was just ABOVE ban..	59%	7%	1%	1%		32%
It was 40% ABOVE £150.	4%	96%				
The price of fuel has jumped ABOV..	4%	93%				2%
The new estimate is 95,000 ABOV...	8%	92%				
Anything ABOVE zero degrees an...	13%	85%			1%	1%
He has a surplus of votes over and...	21%	78%		1%		
Train fares have risen ABOVE inflat...	24%	75%				1%
The shelf is fixed to the wall ABOV...			93%		7%	
The dictionaries are ABOVE the hi...			92%		5%	2%
I've hung some mistletoe ABOVE t...			90%		10%	
There was a faint bruise ABOVE he..			86%		13%	1%
The comments (listed ABOVE) are ...			1%	99%		
As described ABOVE, this uses a n...			1%	99%		
The process, described ABOVE, is ...			1%	99%		
It was refused for the ABOVE reas...			1%	99%		
The ABOVE constraints are not se...			2%	98%		
All of the ABOVE laws have been p..			1%	98%		1%
This site is elevated ABOVE the ro...			25%		75%	
We had a great view from the cliff ...		1%	25%	1%	71%	1%
The stars ABOVE were partly obsc...			29%	1%	70%	
It was built on the hill, just ABOVE ...			33%		66%	1%
Glastonbury Tor towers ABOVE the..			36%		63%	1%
The town is 200m ABOVE sea-leve..		12%	26%		62%	
The plane was cruising ABOVE the...			42%		58%	
We were observed from the windo...			48%	2%	49%	
I used to be his boss, but he works...						100%
ABOVE private soldiers there are t...				1%		99%
There is a level of executives ABO...		1%				99%
The orders came from ABOVE.				1%		99%
We're under serious pressure from...				1%		99%
Now that I work ABOVE her, we do...			3%			97%
I'm ABOVE all that petty business.	37%	1%	1%			60%
They think they're ABOVE work lik...	38%	2%	1%			58%
She's not ABOVE silly gossip.	43%	1%	1%			55%

Below

Popular Placements Matrix

	Worse than	Vantage	Underneath (3D)	Lower than (2D)	Less than	Text uses
He is an unenthusiastic and BELO...	98%				2%	
You wouldn't be doing the job if yo...	82%			2%	17%	
He perfomed BELOW par last time.	80%				20%	
Congress is somewhere BELOW c...	73%		2%		26%	
Paul had performed BELOW expec...	71%				26%	3%
We must set standards of achieve...	70%	2%			27%	2%
The walk provides wonderful view...		98%		2%		
I called out to the people on the be..		97%	2%	2%		
Tabith stood BELOW, watching him.		92%	2%	3%	2%	2%
Your mates are down BELOW, watc..		89%	9%	2%		
The campus was shrinking BELOW...	2%	73%	15%	8%	3%	
They established an iron foundry in..		67%	18%	12%	2%	2%
Instead of being up high the box w...		58%	14%	20%		9%
We dredged BELOW the mud at th...	2%	3%	92%	3%		
The crocodile sank BELOW the sur...		3%	88%	9%		
When we got BELOW the next laye..	5%	5%	83%	6%	2%	
The people in the flat BELOW woul...		42%	53%	5%		
BELOW the crags a well-built tunne..		32%	42%	26%		
I pinned my name badge BELOW t...			2%	94%		5%
She had a mole just BELOW her rig..		2%	2%	94%		3%
There are two iron rings on the wa...		2%	6%	91%		2%
Don't paint BELOW the windowsill.		2%	11%	86%		2%
The sleeves gradually get tighter a...			9%	86%		5%
BELOW the front windows the exte...		17%	2%	82%		
He set a price BELOW the existing ...	3%			2%	95%	
The loss is a little BELOW £3,200.	9%				91%	
We brought in forty million pounds...	9%				91%	
There's no level BELOW which the ...	15%		2%		83%	
There is a £20 surcharge on order...	11%	2%		2%	83%	3%
The sales value was well BELOW t...	20%				80%	
Fill in BELOW all the tasks that you...						100%
Serve with Sharp sauce (see BELO...						100%
Give us your fun verdict by dialling...						100%
In the situations listed BELOW iden...				2%	2%	97%
Look at the sentence BELOW, wha...		2%		2%	2%	95%
It will be argued BELOW that econ...	2%	2%		2%	8%	88%

Appendix 3: Experiment 2 task instructions

It may not be immediately obvious, but many (or even most) words we use have more than one meaning. I am interested in finding out the different meanings of the words *[target words]*. These words are typically understood as having a spatial meaning. A similar word like *in* also has spatial meaning, like this:

1. The ball is **in** the bowl

But it also has other meanings, some of which might be spatial, like 2, or metaphorical, like 3.

2. She's sitting **in** the dark
3. Mary is **in** love with John

The purpose of these examples is to show you that a word as simple as *in* has lots of different meanings; some are obviously related, some take some thinking about before you can see a connection, while some seem completely unrelated to any others.

It is now your task to read through all of the sentences on the left. There are 100 sentences to read. Each of the sentences is a real usage. Read through all of them first, then go back through them and group them on the basis of how the word in capitals is used - the groups should consist of examples in which the word is used in the same way. For example, you might group these examples of *on* together:

The cat sat **on** the table
She's rolling around **on** the floor
Her lunch box is **on** the chair

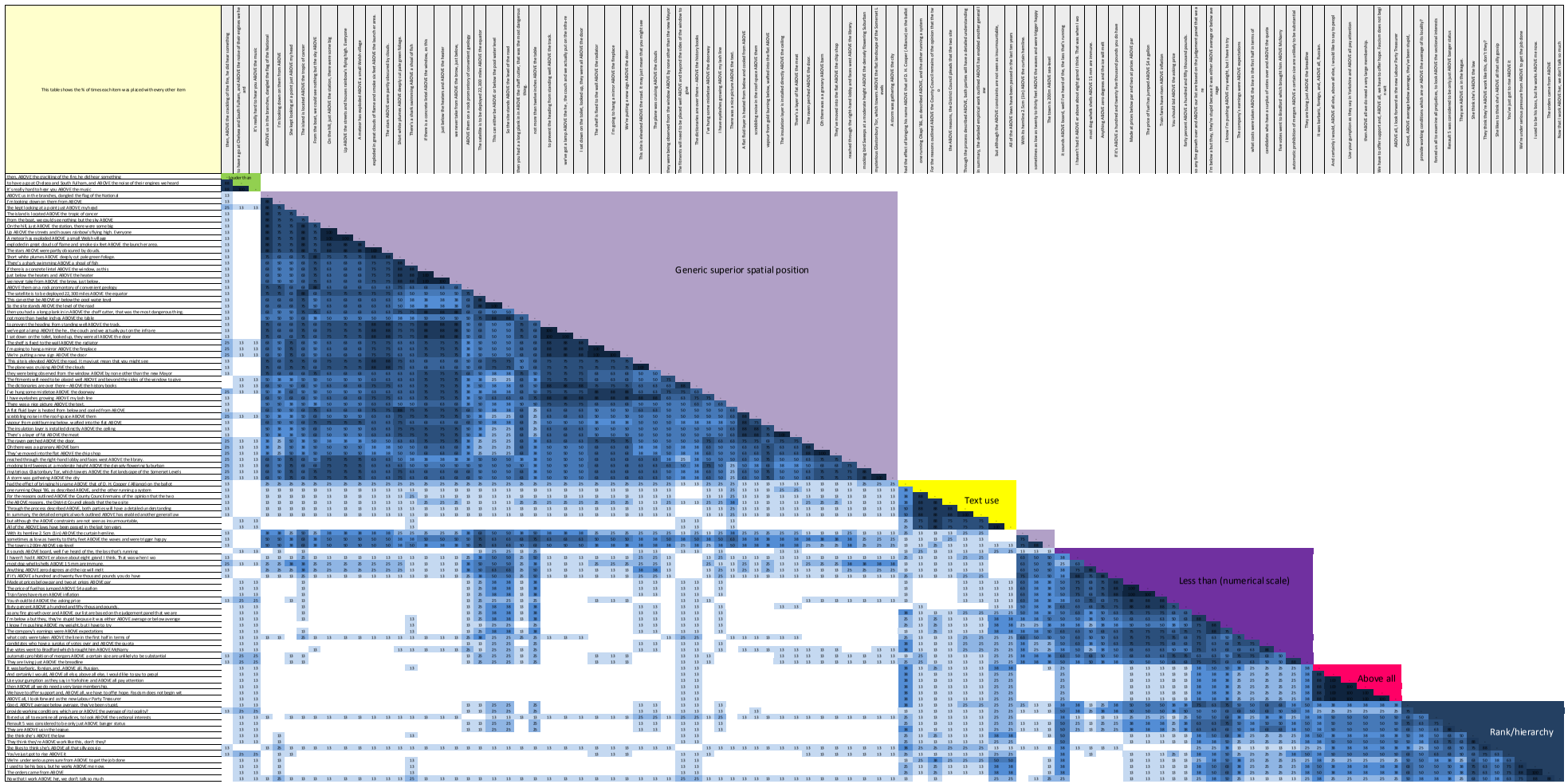
But you might not include examples such as

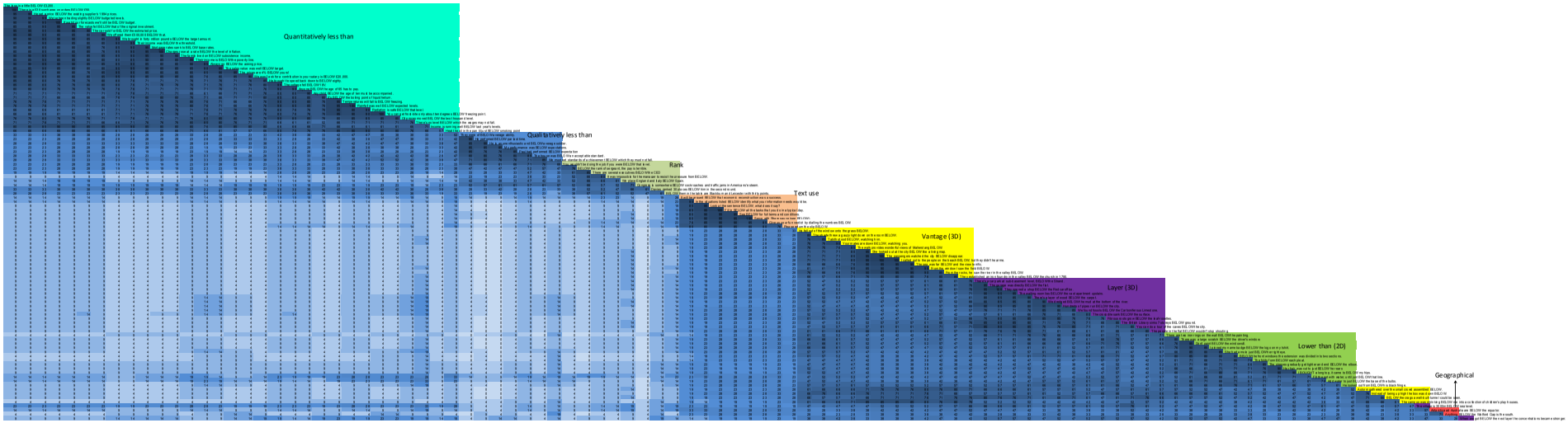
The picture is hanging **on** the wall
She put the lead **on** the dog

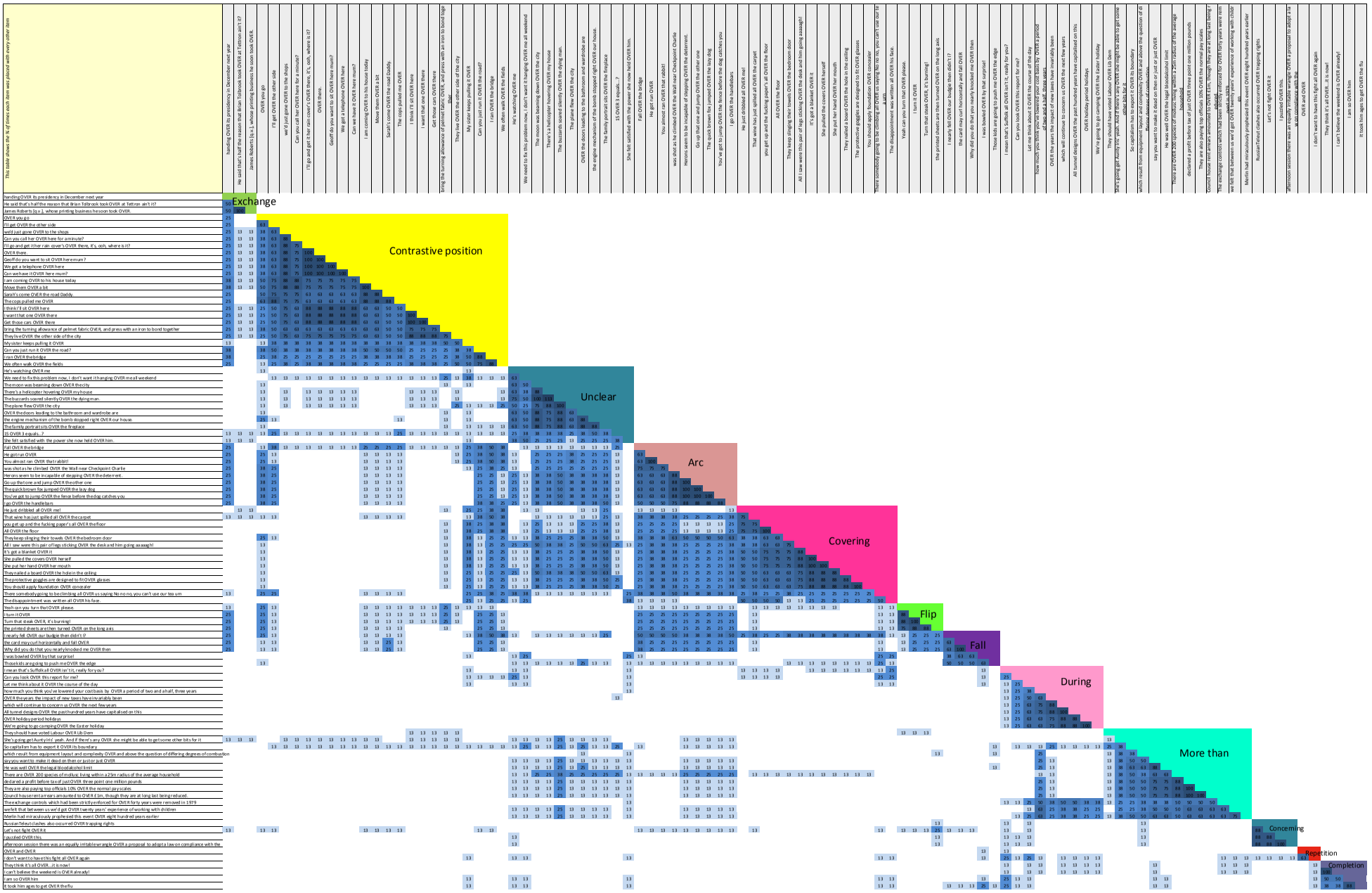
Some of your groups might have lots of examples; some might only have one or two. There is no right answer in this task, I'm just trying to understand what distinctions **you** make when classifying different usages of each word. When you're happy with each group, please write on a piece of card what you think makes each example in that group the same. You might find it easier to say what word/phrase could be used instead of the capitalised word in all cards in the group.

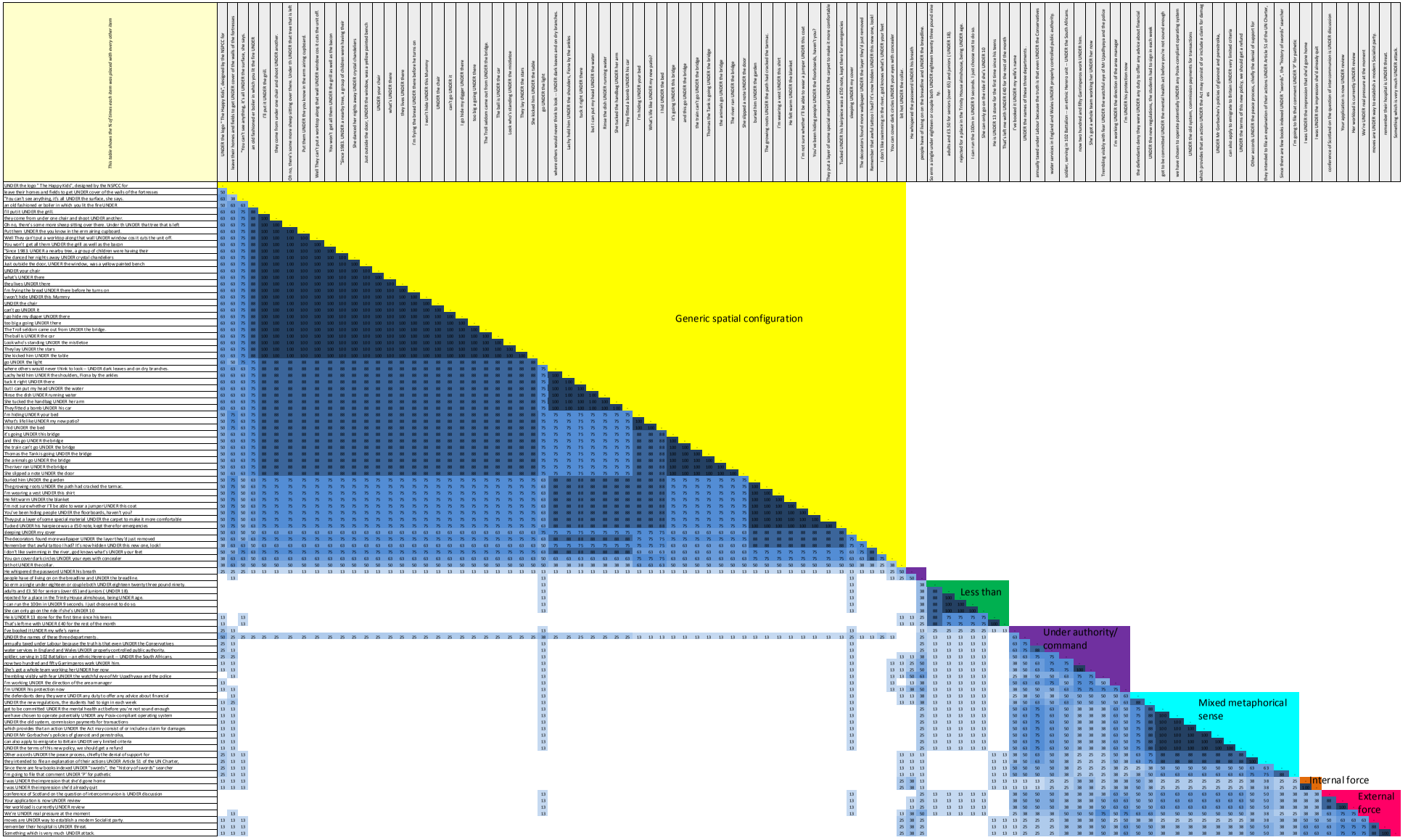
There is no time limit for this task.

Appendix 4: Annotated similarity matrices


Generic superior spatial position
Text use
Less than (numerical scale)
Above all
Rank/hierarchy


Quantitatively less than
Qualitatively less than
Rank
Text use
Vantage (3D)
Layer (3D)
Lower than (2D)
Geographical


This table shows the % of times each item was placed with every other item
handing OVER its presidency in December next year
He said that's half the reason that Brian Tolbrook took OVER at Tetbron ain't it?
James Roberts (q.v.), whose printing business he soon took OVER.
OVER you go
I'll get OVER the other side
we'd just gone OVER to the shops
Can you call her OVER here for a minute?
I'll go and get it her rain cover's OVER there, it's, ooh, where is it?
OVER there.
Geoff do you want to sit OVER here mum?
We got a telephone OVER here
Can we have it OVER here mum?
I am coming OVER to his house today
Move them OVER a bit
Sarah's come OVER the road Daddy.
The cops pulled me OVER
I think I'll sit OVER here
I want that one OVER there
Get those cars OVER there
bring the turning allowance of helmet fabric OVER, and press with an iron to bond together
They live OVER the other side of the city
My sister keeps pulling it OVER
Can you just run it OVER the road?
I can COVER the bridge
We often walk OVER the fields
He's watching OVER me
We need to fix this problem now, I don't want it hanging OVER me all weekend
The moon was beaming down OVER the city
There's a helicopter hovering OVER my house
The buzzards soared silently OVER the dying man.
The plains flew OVER the city
OVER the doors leading to the bathroom and wardrobe are
the engine mechanism of the bomb stopped right OVER our house.
The family portrait sits OVER the fireplace
15 OVER 3 equals...?
She felt satisfied with the power she now held OVER him.
Fell OVER the bridge
He got run OVER
You almost ran OVER that rabbit!
was shot as he climbed OVER the Wall near Checkpoint Charlie
Herons seem to be incapable of stepping OVER the deterrent.
Go up that one and jump OVER the other one.
The quick brown fox jumped OVER the lazy dog.
You've got to jump OVER the fence before the dog catches you
I go OVER the handlebars
He just dribbled all OVER me!
That wine has just spilled all OVER the carpet.
you get up and the fucking paper's all OVER the floor
All OVER the floor.
They keep slinging their towels OVER the bedroom door
All I saw were this pair of legs sticking OVER the desk and him going aaaaagh!
It's got a blanket OVER it
She pulled the covers OVER herself
She put her hand OVER her mouth
They nailed a board OVER the hole in the ceiling
The protective goggles are designed to fit OVER glasses
You should apply foundation OVER concealer
There somebody going to be climbing all OVER us saying No no no, you can't use our tea urn
The disappointment was written all OVER his face.
Yeah can you turn that OVER please.
I turn it OVER
Turn that steak OVER, it's burning!
the printed sheets are then turned OVER on the long axis.
I nearly fell OVER our budgie then didn't I?
the card may curl horizontally and fall OVER.
Why did you do that you nearly knocked me OVER then
I was bowled OVER by that surprise!
Those kids are going to push me OVER the edge
I mean that's Suffolk all OVER isn't it, really for you?
Can you look OVER this report for me?
Let me think about it OVER the course of the day.
how much you think you've lowered your cost basis by OVER a period of two and a half, three years
OVER the years the impact of new laws have invariably been
which will continue to concern us OVER the next few years
All tunnel designs OVER the past hundred years have capitalised on this
OVER holiday period holidays
We're going to go camping OVER the Easter holiday
They should have voted Labour OVER Lib Dem
She's going get Aunty Int. yeah. And if there's any OVER she might be able to get some other bits for it
So capitalism has to export it OVER its boundary
which result from equipment layout and complexity OVER and above the question of differing degrees of combustion
as you want to, make it dead on then or just or just OVER
He was well OVER the legal bloodalcohol limit.
There are OVER 200 species of mollusc living within a 25m radius of the average household
declared a profit before tax of just OVER three point one million pounds
They are also paying top officials 10% OVER the normal pay scales.
Council house rent arrears amounted to OVER £3m, though they are at long last being reduced.
The exchange controls which had been strictly enforced for OVER forty years were removed in 1979
we felt that between us we'd got OVER twenty years' experience of working with children
Merlin had miraculously prophesied this event OVER eight hundred years earlier.
Russian Telcot clashes also occurred OVER trapping rights
Let's not fight OVER it
I jaunted OVER this.
afternoon session there was an equally irritable wrangle OVER a proposal to adopt a law on compliance with the OVER and OVER
I don't want to have this fight all OVER again
They think it's all OVER...it is now!
I can't believe the weekend is OVER already!
I am so OVER him
It took him ages to get OVER the flu
Exchange
Contrastive position
Unclear
Arc
Covering
Flip
Fall
During
More than
Concerning
Repetition
Completion


Generic spatial configuration
Less than
Under authority/ command
Mixed metaphorical sense
Internal force
External force

Appendix 5: Experiment 3 stimuli

Mixed sentences	Spatial sentences	Non-spatial sentences
I'm wearing a vest UNDER this shirt.	I'm wearing a vest UNDER this shirt.	You can emigrate to Britain UNDER limited criteria.
Tracksuit bottoms should not be worn UNDER school trousers.	Tracksuit bottoms should not be worn UNDER school trousers.	He will be committed UNDER the mental health act.
Should I wear a jumper UNDER this coat?	Should I wear a jumper UNDER this coat?	It will be contested UNDER the Corrupt Practices Act.
They got UNDER cover of the walls of the fortresses.	They got UNDER cover of the walls of the fortresses.	They will file an explanation UNDER Article 51 of the UN Charter.
Rinse the dish UNDER running water.	Rinse the dish UNDER running water.	UNDER the new regulations, the students must sign in each week.
They found more wallpaper UNDER the layer they'd removed.	They found more wallpaper UNDER the layer they'd removed.	The fridge can be exchanged UNDER the returns policy.
You can emigrate to Britain UNDER limited criteria.	I'm frying the bread UNDER there.	Remember their hospital is UNDER threat.
He will be committed UNDER the mental health act.	I'm hiding UNDER your bed.	This is something which is very much UNDER attack.
It will be contested UNDER the Corrupt Practices Act.	The Troll seldom came out from UNDER the bridge.	They deny they were UNDER any duty to offer any advice.
They will file an explanation UNDER Article 51 of the UN Charter.	She kicked him UNDER the table.	We're UNDER real pressure at the moment.
UNDER the new regulations, the students must sign in each week.	They lay UNDER the stars.	Your application is now UNDER review.
The fridge can be exchanged UNDER the returns policy.	Cover dark circles UNDER your eyes with concealer.	The question of intercommunion is UNDER discussion.
Remember their hospital is UNDER threat.	I'm sleeping UNDER my cover.	We'll be taxed UNDER the Conservatives.
This is something which is very much UNDER attack.	The growing roots UNDER the path had cracked the tarmac.	I'm working UNDER the direction of the area manager.
They deny they were UNDER any duty to offer any advice.	They put material UNDER the carpet to make it more comfortable.	250 Garrimperos work UNDER him.
We're UNDER real pressure at the moment.	They looked where others wouldn't think to - UNDER dark leaves.	She's got a whole team working UNDER her now.
Your application is now UNDER review.	They'd been hiding people UNDER the floorboards.	He served in 102 Battalion UNDER the South Africans.
The question of intercommunion is UNDER discussion.	He felt warm UNDER the blanket.	The Act was introduced UNDER the last President.
I'm frying the bread UNDER there.	The creatures must adapt to life UNDER ice-covered water.	Alabama has a population of just UNDER 10,000 in 1810.
I'm hiding UNDER your bed.	Barium metal is kept UNDER mineral oil to prevent a reaction.	Armstrong and Aldrin returned it just UNDER a day later.
The Troll seldom came out from UNDER the bridge.	One third of the town was UNDER water.	3.1% households has children UNDER 18 living with them.
She kicked him UNDER the table.	The U-boats will remain UNDER the sea.	I brought the project in UNDER budget at £108,065
They lay UNDER the stars.	The Apollo film collection had been stored UNDER liquid nitrogen.	She has just under 1 billion to spend.
Cover dark circles UNDER your eyes with concealer.	Harry rescued Ron from UNDER the lake.	Most salamanders are UNDER 15cm long.
I'm sleeping UNDER my cover.	The fabric extends from the ankle to UNDER the foot.	Gray published most of her work UNDER a pseudonym.
The growing roots UNDER the path had cracked the tarmac.	The saddle slipped UNDER the horse's belly.	He enrolled UNDER the guise of Bobby Parker.
They put material UNDER the carpet to make it more comfortable.	Nylon straps pass UNDER the hull of the ship.	My account is UNDER my wife's name.
They looked where others wouldn't think to - UNDER dark leaves.	Neon lights UNDER the car give the car a distinctive appearance.	The company produced a version UNDER the name "Snatch".
They'd been hiding people UNDER the floorboards.	The rider's helmet was tied UNDER her chin.	He was working UNDER the assumed name of C. Lenning.
He felt warm UNDER the blanket.	There were no stickers UNDER the bottom of the skateboard.	He attended hospital UNDER the name of Graham Ferguson.
We'll be taxed UNDER the Conservatives.	Only small boats can pass UNDER the bridge.	She was working UNDER the assumption she'd be fired.
I'm working UNDER the direction of the area manager.	The note had been pushed UNDER the door.	We were UNDER the impression the play was cancelled.
250 Garrimperos work UNDER him.	The river passes UNDER the Tunnel Avenue Bridge.	The Vice-Chancellor was there UNDER a misapprehension.
She's got a whole team working UNDER her now.	The subway line runs UNDER the overpass.	UNDER this reasoning, the court was right to affirm this verdict.
He served in 102 Battalion UNDER the South Africans.	When the sun sets it does not pass UNDER the Earth.	We tolerated his behaviour UNDER the belief that he was truly gifted.
The Act was introduced UNDER the last President.	The players were forced to drive UNDER the bridge.	The people boarded the plane UNDER false pretenses.

Mixed sentences	Spatial sentences	Non-spatial sentences
There was a wrangle OVER a proposal to adopt a law.	Can you just run it OVER the road?	There was a wrangle OVER a proposal to adopt a law.
I puzzled OVER this.	The cops pulled me OVER.	I puzzled OVER this.
Let's not fight OVER it.	I ran OVER the bridge.	Let's not fight OVER it.
Clashes also occurred OVER trapping rights.	Sarah's come OVER the road Daddy.	Clashes also occurred OVER trapping rights.
We had some discussions OVER where to place the boundaries.	The plane flew OVER the city.	We had some discussions OVER where to place the boundaries.
We'd fall out OVER stupid things and not speak to each other.	He walked slowly OVER the zebra-crossing.	We'd fall out OVER stupid things and not speak to each other.
Can you just run it OVER the road?	Jump OVER the other one.	He is handing OVER his presidency.
The cops pulled me OVER.	Herons seem to be incapable of stepping OVER the deterrent.	That's half the reason that Brian Tolbrook took OVER at Tettron.
I ran OVER the bridge.	I go OVER the handlebars.	He took OVER the printing business.
Sarah's come OVER the road Daddy.	The quick brown fox jumped OVER the lazy dog.	I'll take OVER the primary agenda.
The plane flew OVER the city.	He refused to return the balls kicked OVER his fence.	I can't hand OVER a long barrelled weapon to that officer.
He walked slowly OVER the zebra-crossing.	They keep slinging their towels OVER the bedroom door.	The plaintiff handed OVER to Samuel Revill the first note.
Jump OVER the other one.	Bring pelmet fabric OVER, and press with an iron to bond together.	I rushed out before the show was OVER.
Herons seem to be incapable of stepping OVER the deterrent.	I turn it OVER.	I can't believe the weekend is OVER already!
I go OVER the handlebars.	The printed sheets are turned OVER on the long axis.	His wrestling days are not OVER yet.
The quick brown fox jumped OVER the lazy dog.	Turn that steak OVER, it's burning!	They think it's all OVER…it is now!
He refused to return the balls kicked OVER his fence.	Yeah can you turn that OVER please.	She wished the party was OVER.
They keep slinging their towels OVER the bedroom door.	He saw a car flip OVER and land upside down in a hedge.	We hope Sunderland will go up once the game is OVER.
He is handing OVER his presidency.	Jake watched as the pins tumbled OVER.	OVER the next 17 months, he would be imprisoned in the camps.
That's half the reason that Brian Tolbrook took OVER at Tettron.	I would always fall OVER as a kid.	The ratios have varied OVER the centuries.
He took OVER the printing business.	It takes at least two people to push OVER a cow.	OVER its history, anthropology has developed a number of related concepts.
I'll take OVER the primary agenda.	The world contains physics to make blocks topple OVER.	Lizards have been increasingly found in homes as pets OVER the last few years.
I can't hand OVER a long barrelled weapon to that officer.	An explosion caused the tree to fall OVER.	The economy would continue to decline OVER a 24-month period.
The plaintiff handed OVER to Samuel Revill the first note.	Equipment on the set had been accidentally knocked OVER.	Preparation for the event extended OVER the years 1859-1866.
I rushed out before the show was OVER.	This oil-producing region is spread OVER much of western Pennsylvania.	Research showed impacts of OVER 10%.
I can't believe the weekend is OVER already!	The growth develops OVER the whole agar surface.	Union enlistments increased by OVER 10,000.
His wrestling days are not OVER yet.	The ranch extends OVER much of Pecos County.	Amazon has OVER 900,000 members in its affiliate programme.
They think it's all OVER…it is now!	The painting stretches OVER 500sq m of ceiling.	Factory workers are paid just OVER $12 per hour.
She wished the party was OVER.	The new design was actually spread OVER two stamps.	Pennies from Heaven has raised well OVER $2 million.
We hope Sunderland will go up once the game is OVER.	A piece of paper was pasted OVER Somerset's name in the report.	The protest cost taxpayers OVER a million pounds.
Bring pelmet fabric OVER, and press with an iron to bond together.	Ornate carvings are found OVER the doorway.	He tried to increase his control OVER the landowners.
I turn it OVER.	A painting hangs OVER the mantel of the Roosevelt Room.	The commission had little power OVER networks.
The printed sheets are turned OVER on the long axis.	He played the banjo with his hands OVER his head.	He presided OVER the lunar landings.
Turn that steak OVER, it's burning!	The inscriptions were fixed OVER the gateway of the larger courtyard.	Mr Bush presided OVER the first meeting of his new team.
Yeah can you turn that OVER please.	OVER the desk hangs a portrait of Anton Rubenstein.	There were manifestations of Roman authority OVER other churches.
He saw a car flip OVER and land upside down in a hedge.	Drones were spotted hovering OVER the bomb site.	Marriage gave a couple rights OVER each other.

Mixed sentences	Spatial sentences	Non-spatial sentences
Congress is somewhere BELOW cockroaches and traffic jams in Americans' esteem.	Tabith stood BELOW, watching him.	Congress is somewhere BELOW cockroaches and traffic jams in Americans' esteem.
We must set standards of achievement BELOW which they must not fall.	BELOW the front windows the extension was divided into two sections.	We must set standards of achievement BELOW which they must not fall.
You wouldn't be doing the job if you were BELOW that level.	They established an iron foundry in the valley BELOW the church in 1790.	You wouldn't be doing the job if you were BELOW that level.
Paul had performed BELOW expectation.	The walk provides wonderful views of Mallerstang BELOW.	Paul had performed BELOW expectation.
He performed BELOW par last time.	Your mates are down BELOW, watching you.	He performed BELOW par last time.
He is an unenthusiastic and BELOW average soldier.	I called out to the people on the beach BELOW, but they didn't hear me.	He is an unenthusiastic and BELOW average soldier.
Tabith stood BELOW, watching him.	The people in the flat BELOW wouldn't stop shouting.	The loss is a little BELOW £3,200.
BELOW the front windows the extension was divided into two sections.	Instead of being up high the box was down BELOW.	There's no level BELOW which the wages may not fall.
They established an iron foundry in the valley BELOW the church in 1790.	We dredged BELOW the mud at the bottom of the river.	We brought in forty million pounds BELOW the target amount.
The walk provides wonderful views of Mallerstand BELOW.	When we got BELOW the next layer the concentrations became stronger.	He set a price BELOW the existing supplier's 1994 prices.
Your mates are down BELOW, watching you.	The campus was shrinking BELOW me into a collection of children's play houses.	There is a £20 surcharge on orders BELOW £50.
I called out to the people on the beach BELOW, but they didn't hear me.	The crocodile sank BELOW the surface.	The sales value was well BELOW target.
The people in the flat BELOW wouldn't stop shouting.	She had a mole just BELOW her right eye.	The city experiences more than 60 days BELOW 0 degrees C.
Instead of being up high the box was down BELOW.	Don't paint BELOW the windowsill.	The final temperature is BELOW the starting value.
We dredged BELOW the mud at the bottom of the river.	BELOW the crags a well-built tunnel could be seen.	Liquids may change to vapour at temperatures BELOW their boiling points.
When we got BELOW the next layer the concentrations became stronger.	I pinned my name badge BELOW the logo on my tshirt.	It is possible to supercool liquid water to BELOW the freezing point.
The campus was shrinking BELOW me into a collection of children's play houses.	There are two iron rings on the wall BELOW the painting.	Minimum temperatures rarely fall BELOW 38 C.
The crocodile sank BELOW the surface.	The sleeves gradually get tighter and end BELOW the elbow.	Temperatures range from BELOW -2 to over 30 C.
She had a mole just BELOW her right eye.	Fill in BELOW all the tasks that you do in a typical day.	BELOW him there was an administrative hierarchy.
Don't paint BELOW the windowsill.	Look at the sentence BELOW, what does it say?	BELOW the nobility were priests, physicians and engineers.
BELOW the crags a well-built tunnel could be seen.	Give us your fun verdict by dialling the numbers BELOW.	Those BELOW him in the feudal chain owed nothing to Frederick.
I pinned my name badge BELOW the logo on my tshirt.	Serve with Sharp sauce (see BELOW).	The principal relied on the support of those BELOW her.
There are two iron rings on the wall BELOW the painting.	In the situations listed BELOW identify what your information needs would be.	The chief eunuch would direct junior eunuchs BELOW him.
The sleeves gradually get tighter and end BELOW the elbow.	It will be argued BELOW that economic reconstruction was a success.	BELOW the leader are numerous clan or family chiefs.
The loss is a little BELOW £3,200.	The bird is dark rusty-brown above and dark grey BELOW.	Tara Brown is in the year BELOW the four main characters.
There's no level BELOW which the wages may not fall.	Adults are light olive above, yellow BELOW, and have a black hood.	Bob was in the year BELOW.
We brought in forty million pounds BELOW the target amount.	The female mandarin duck is paler BELOW.	Griff Rhys Jones was a year BELOW him, and in the same class as Charles Thomson.
He set a price BELOW the existing supplier's 1994 prices.	Wheatears are blue-grey above and white BELOW.	They studied at Stockbridge High and were a year BELOW the rest of their teammates.
There is a £20 surcharge on orders BELOW £50.	Nightingales are generally buff to white BELOW.	You'd never consider going out with a boy from the year BELOW.
The sales value was well BELOW target.	They are dark green and shiny above, and rusty and hairy BELOW.	Gee was in the year BELOW Princess Anne.
Fill in BELOW all the tasks that you do in a typical day.	The house stood 300 yards BELOW the county line.	It was BELOW a Gentlewoman to wrangle.
Look at the sentence BELOW, what does it say?	Antarctica and Australia are BELOW the equator.	It is too far BELOW contempt to be worth castigating.
Give us your fun verdict by dialling the numbers BELOW.	Anything BELOW the Watford Gap is the south.	It is BELOW her to say that.
Serve with Sharp sauce (see BELOW).	My daughter once went to the Cook Islands BELOW Hawaii.	I am BELOW that kind of work if nothing else.
In the situations listed BELOW identify what your information needs would be.	Is South America above or BELOW the Tropic of Cancer?	Such petty behavior is BELOW me.
It will be argued BELOW that economic reconstruction was a success.	Fifteen miles BELOW the border, you'll face a checkpoint.	Good students thought it was BELOW them, and not challenging enough.

Mixed sentences	Spatial sentences	Non-spatial sentences
They think they're ABOVE work like this.	There was a faint bruise ABOVE her eyebrow.	They think they're ABOVE work like this.
Are they good, ABOVE average, or below average?	I've hung some mistletoe ABOVE the doorway.	Are they good, ABOVE average, or below average?
It was either ABOVE average or below average.	The dictionaries are ABOVE the history books.	It was either ABOVE average or below average.
The Renault 5 was just ABOVE banger status.	She took the form of a mermaid ABOVE the waist.	The Renault 5 was just ABOVE banger status.
She's not ABOVE silly gossip.	The shelf is fixed to the wall ABOVE the radiator.	She's not ABOVE silly gossip.
I'm ABOVE all that petty business.	James had an X-shaped scar ABOVE his nose.	I'm ABOVE all that petty business.
Anything ABOVE zero degrees and the ice will melt	The comments (listed ABOVE) are worrying.	Anything ABOVE zero degrees and the ice will melt
He has a surplus of votes over and ABOVE the quota.	As described ABOVE, this uses a new operating system.	He has a surplus of votes over and ABOVE the quota.
It was 40% ABOVE £150.	It was refused for the ABOVE reasons.	It was 40% ABOVE £150.
The new estimate is 95,000 ABOVE the original estimate.	The process, described ABOVE, is clear to all.	The new estimate is 95,000 ABOVE the original estimate.
The price of fuel has jumped ABOVE $4 a gallon.	All of the ABOVE laws have been passed in the last ten years.	The price of fuel has jumped ABOVE $4 a gallon.
Train fares have risen ABOVE inflation.	The ABOVE constraints are not seen as insurmountable.	Train fares have risen ABOVE inflation.
There was a faint bruise ABOVE her eyebrow.	This site is elevated ABOVE the road.	I used to be his boss, but he works ABOVE me now.
I've hung some mistletoe ABOVE the doorway.	Glastonbury Tor towers ABOVE the Somerset Levels.	Now that I work ABOVE her, we don't talk so much.
The dictionaries are ABOVE the history books.	It was built on the hill, just ABOVE the station.	The orders came from ABOVE.
The plane was cruising ABOVE the clouds.	We had a great view from the cliff ABOVE the cove.	There is a level of executives ABOVE the vice president level.
The shelf is fixed to the wall ABOVE the radiator.	The town is 200m ABOVE sea-level.	ABOVE private soldiers there are three types of officer.
The stars ABOVE were partly obscured by clouds.	We were observed from the window ABOVE.	We're under serious pressure from ABOVE to get the job done.
The comments (listed ABOVE) are worrying.	The male and female common poorwill are similar, both gray and black ABOVE.	I heard their screaming faintly ABOVE the roar of the Martian's collapse.
As described ABOVE, this uses a new operating system.	The bird is dark rusty-brown ABOVE and dark grey below.	Tyrone shouted ABOVE the noise.
It was refused for the ABOVE reasons.	They are dark green and shiny ABOVE, and rusty and hairy below.	You may hear them ABOVE the shouts of battle.
The process, described ABOVE, is clear to all.	Adults are mainly golden-olive ABOVE with buff spots on the wings.	Sometimes you have to listen for it ABOVE the yell of the crowd.
All of the ABOVE laws have been passed in the last ten years.	Their plumage is green and yellow ABOVE and yellow below.	It was hard to hear anything ABOVE the din of the restaurant.
The ABOVE constraints are not seen as insurmountable.	Females and juveniles are brown ABOVE with brown barring below.	I couldn't hear myself think ABOVE the guys' chatter.
This site is elevated ABOVE the road.	There is a cluster of coral pileups just ABOVE the equator.	He was placed a year ABOVE others of his age at school.
Glastonbury Tor towers ABOVE the Somerset Levels.	They were all along up ABOVE the Arctic Circle.	The building houses all students from 9th grade and ABOVE.
It was built on the hill, just ABOVE the station.	The town of Glen Dale is ABOVE the Mason-Dixon line.	Each child is assigned a 'mum' or 'dad' in the year ABOVE.
We had a great view from the cliff ABOVE the cove.	They travelled to the islets ABOVE Hawaii.	Singh was in the year group team two years ABOVE his own.
The town is 200m ABOVE sea-level.	The ship made regular runs across the Artic Circle ABOVE the North Sea.	Entry to the school in year 1 and ABOVE begins with a 'Taster Day'.
We were observed from the window ABOVE.	The plant's distribution extends just ABOVE the border into Arizona.	A girl from the year ABOVE me got a VW bug.
I used to be his boss, but he works ABOVE me now.	The stars ABOVE were partly obscured by clouds.	I think the quality I prize ABOVE all others is curiosity.
Now that I work ABOVE her, we don't talk so much.	The plane was cruising ABOVE the clouds.	She valued reason ABOVE any other human virtue.
The orders came from ABOVE.	This model was designed to attack enemy aircraft up to 5km ABOVE the launching aircraft.	He desired ABOVE all else to restore the political heritage of his family.
There is a level of executives ABOVE the vice president level.	They were mercilessly bombed from ABOVE by the Italian air force.	He felt ABOVE anything that his work was an evolving process.
ABOVE private soldiers there are three types of officer.	It positions itself on a tree limb high ABOVE the forest floor.	She desperately worships beauty ABOVE anything else.
We're under serious pressure from ABOVE to get the job done.	The alien appeared to move through the chamber ABOVE the audience.	"Apple of my eye" refers to someone that one cherishes ABOVE all others.

Appendix 6: Experiment 3 task instructions

Changes to the OptimalSort system mean that, at the time this experiment was undertaken, only short instructions could be presented.

Your task is to organise a list of sentences into groups according to the meaning of the word [target word]. The goal is to end up with groups in which the meaning of [target word] is the same in all sentences in the group. Grouping them is easy:

Step 1

Read through all of the sentences and think about what [target word] means in each one.

Step 2

Drag a sentence into the blank space on the right to create a group.

Step 3

Drag other sentences into that group **so that the meaning of [target word] is the same in all of the sentences in the group.**

Make more groups by dropping sentences in unused spaces.

Step 4

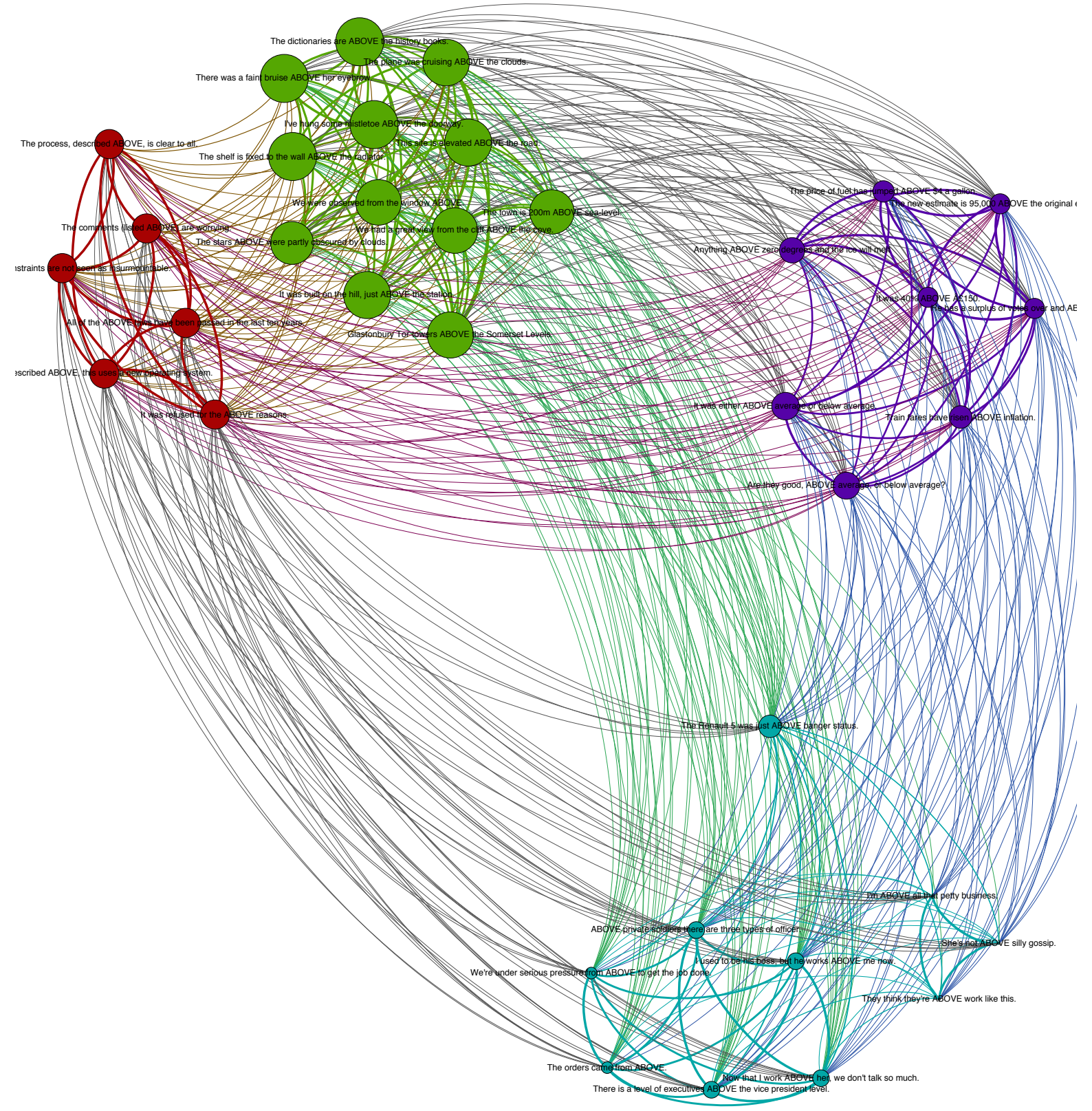
Give each group a name that describes the meaning of [target word] in that group.

Create as many or as few groups as you need, with as many or as few members as you wish. You can move sentences into other groups if you change your mind.

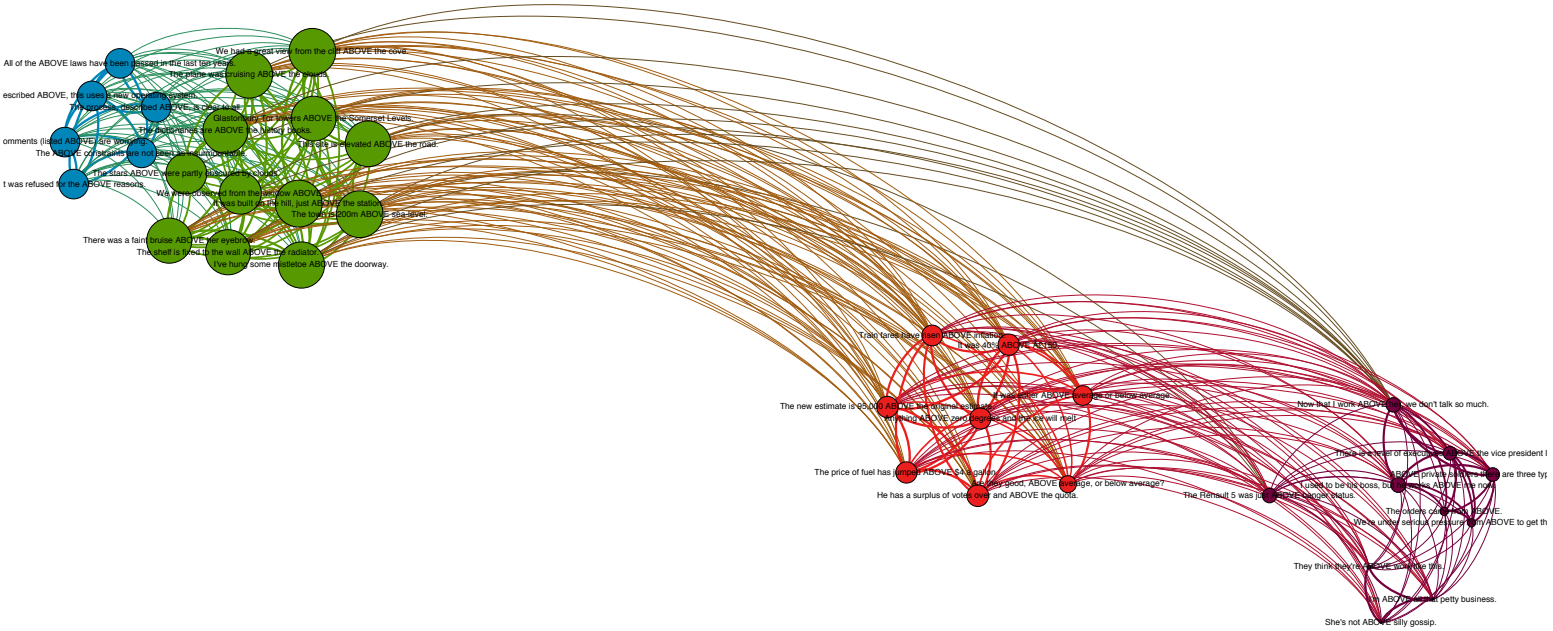
When you're done click "Finished" at the top right. Have fun!

Appendix 7: Network visualisations of Experiment 3 sentence-sorting data

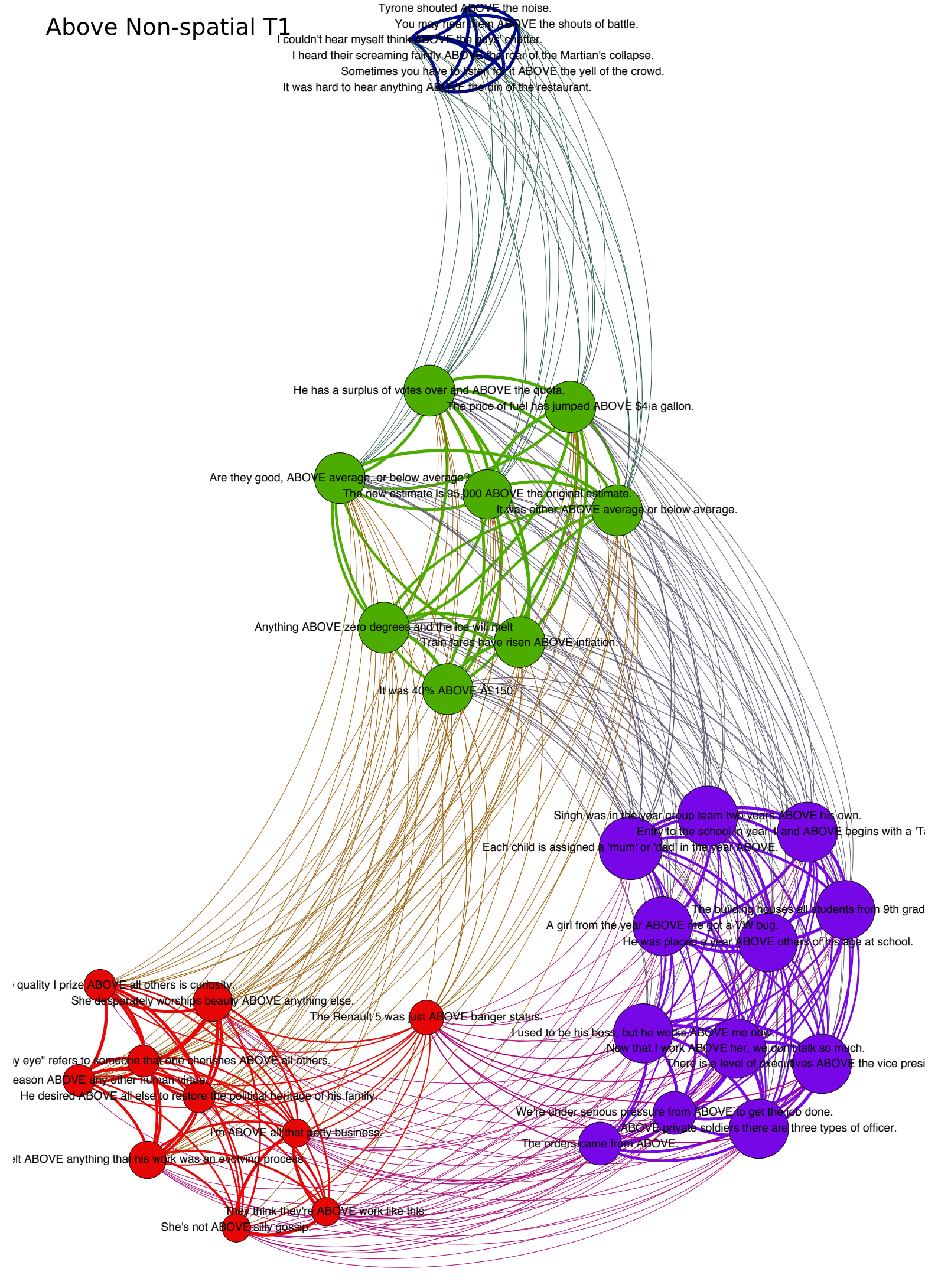
Above Mixed T1



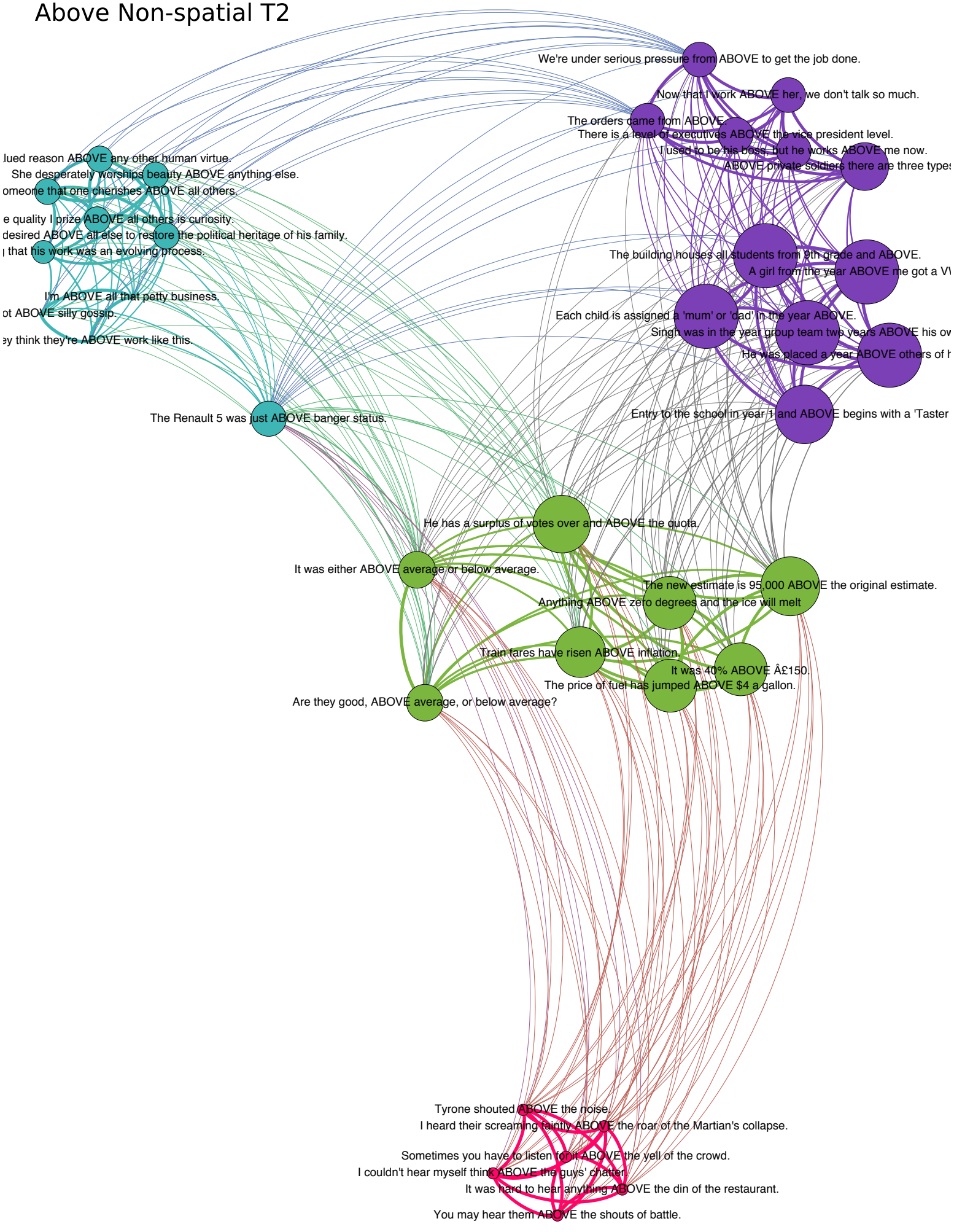
Above Mixed T2



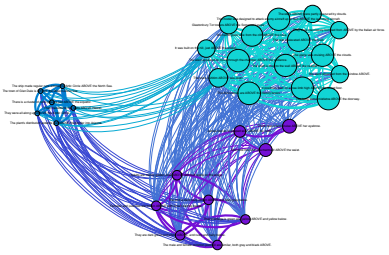
Above Non-spatial T1



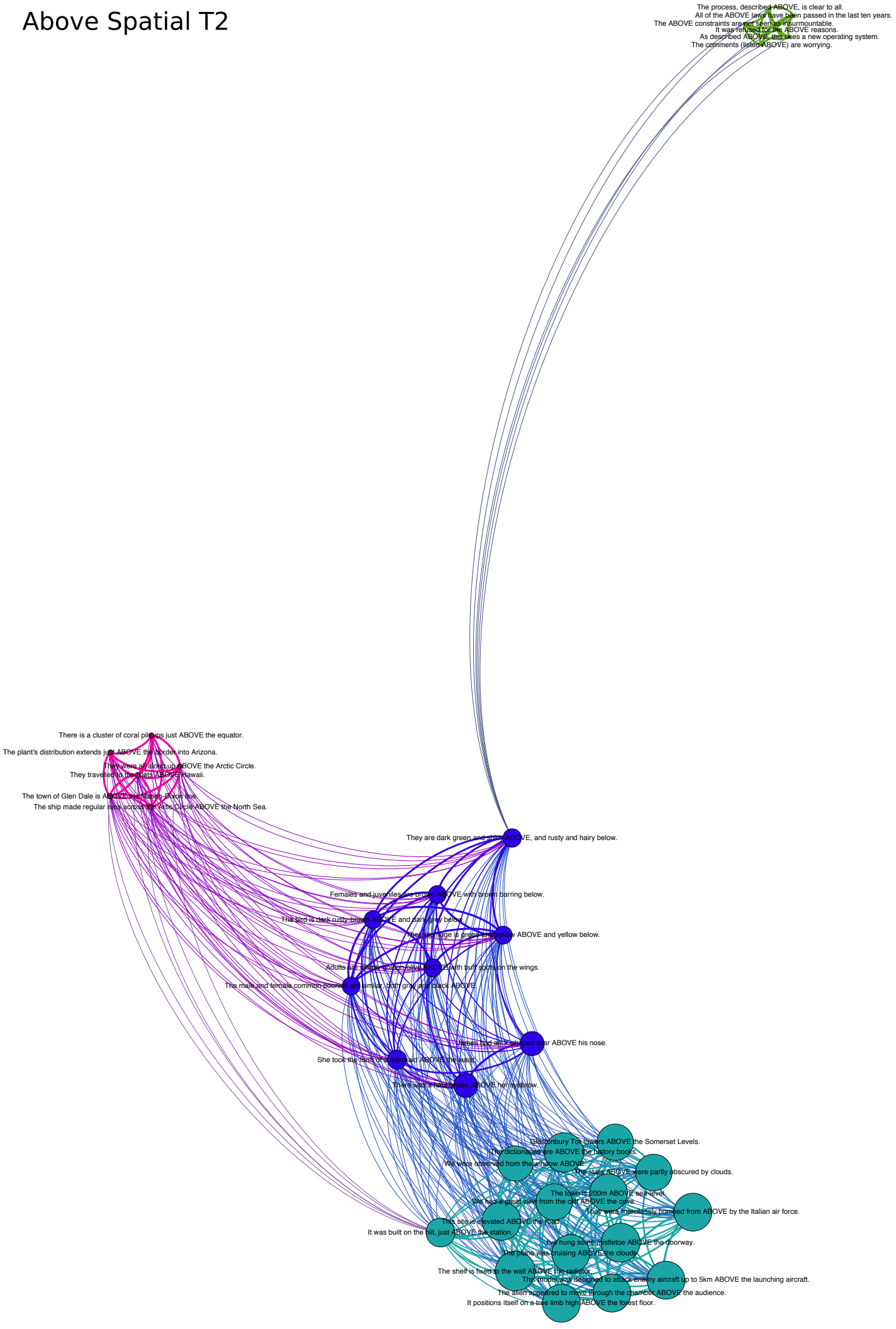
Above Non-spatial T2



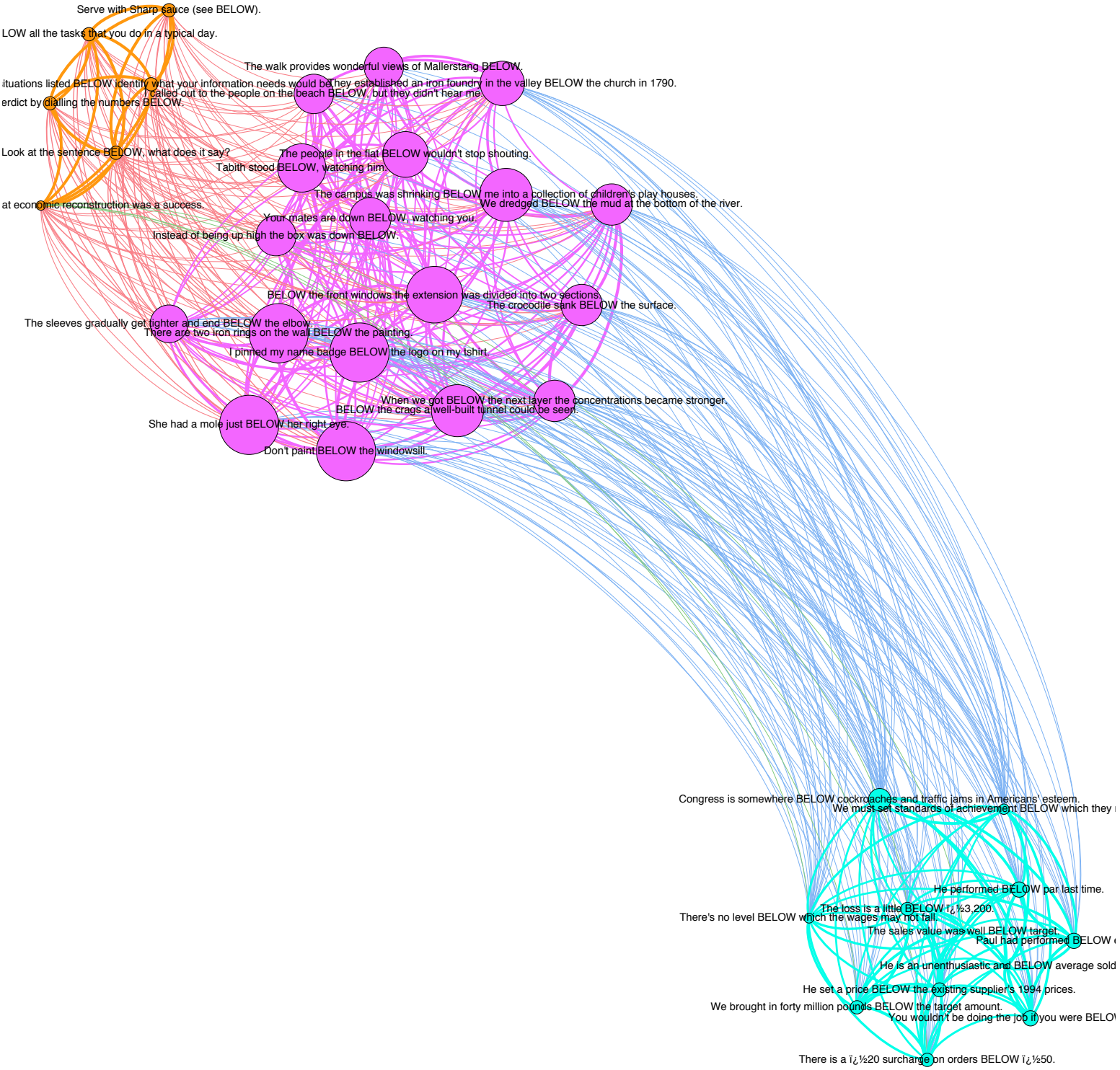
Above Spatial T1



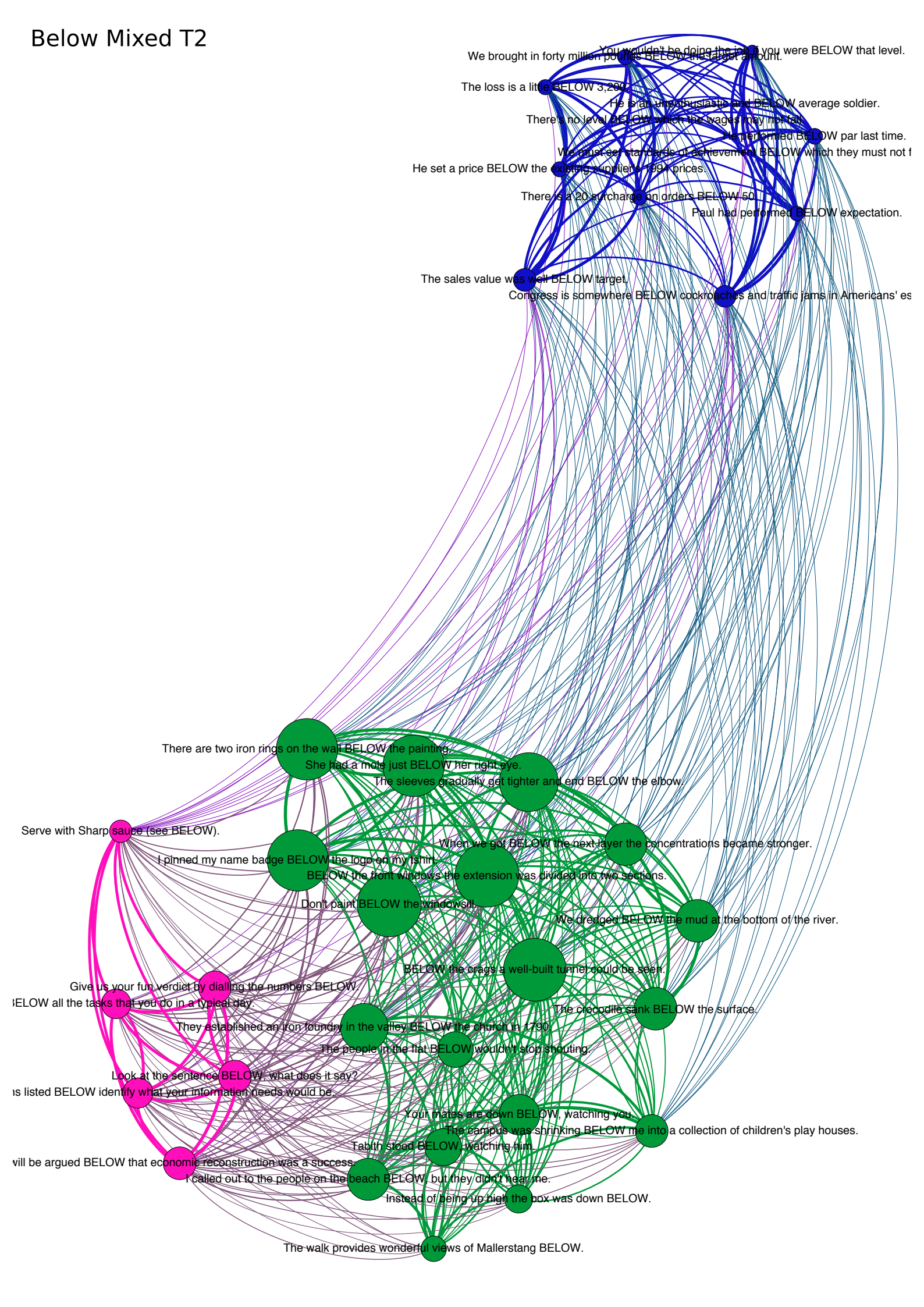
Above Spatial T2



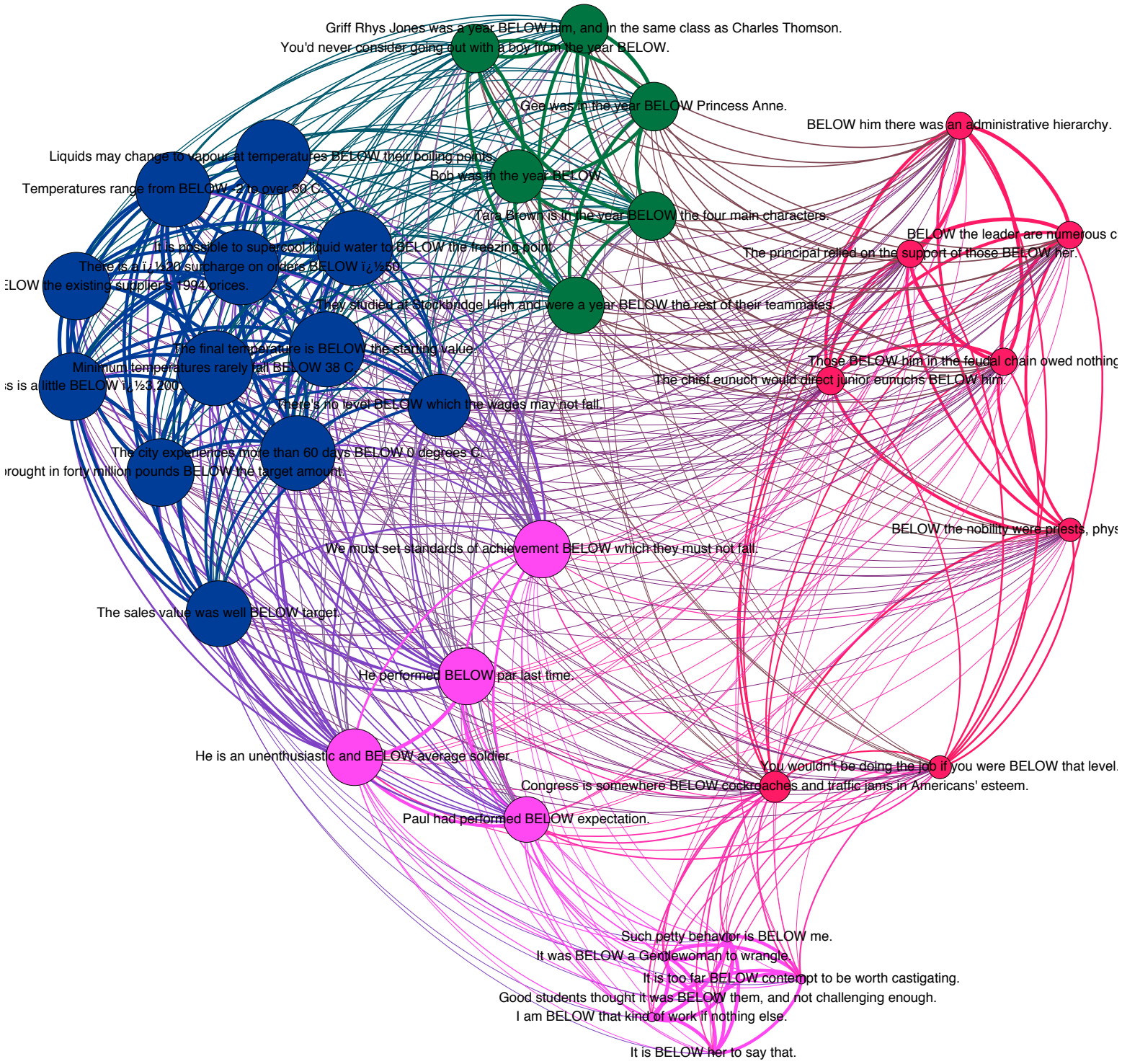
Below Mixed T1



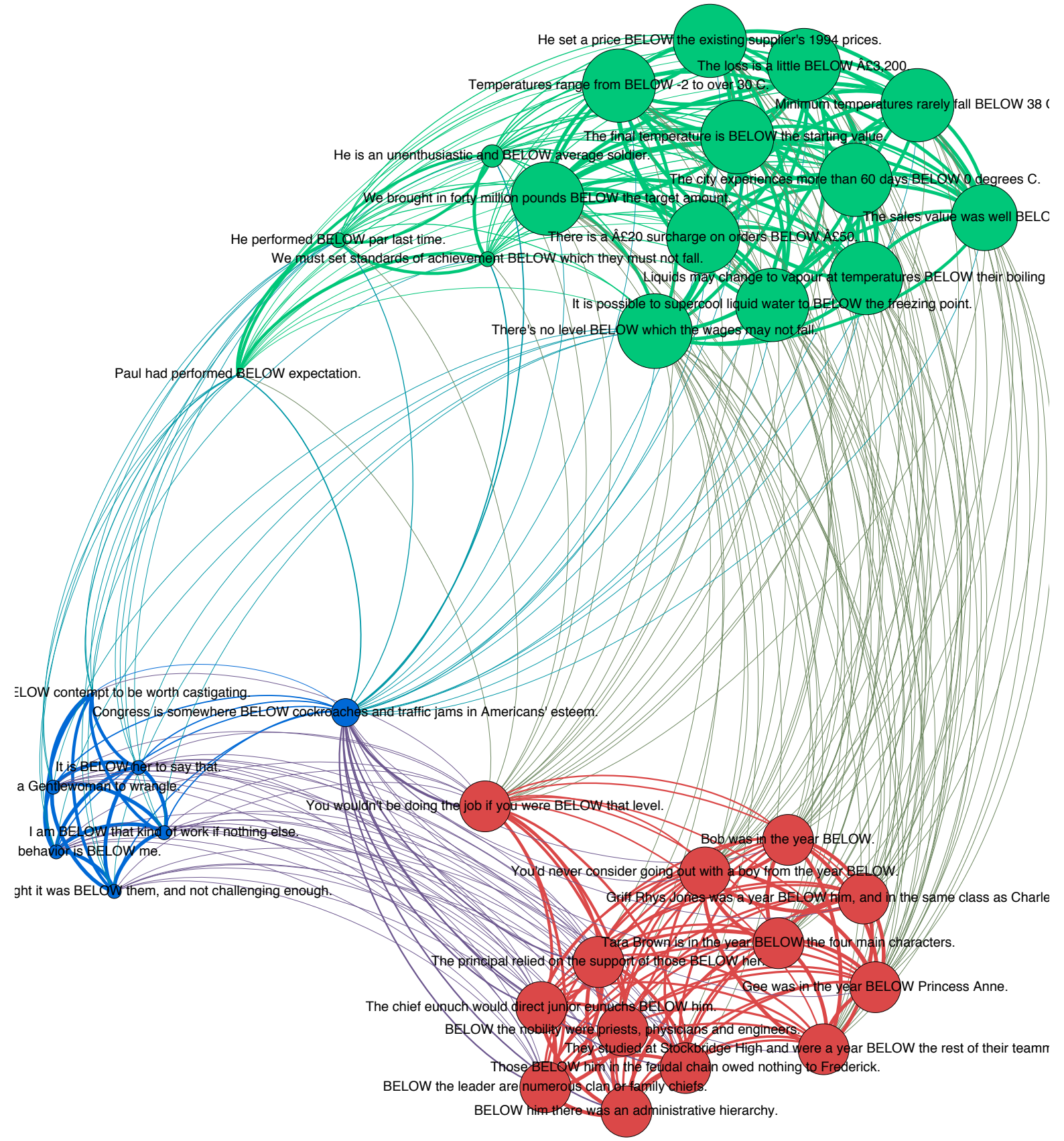
Below Mixed T2



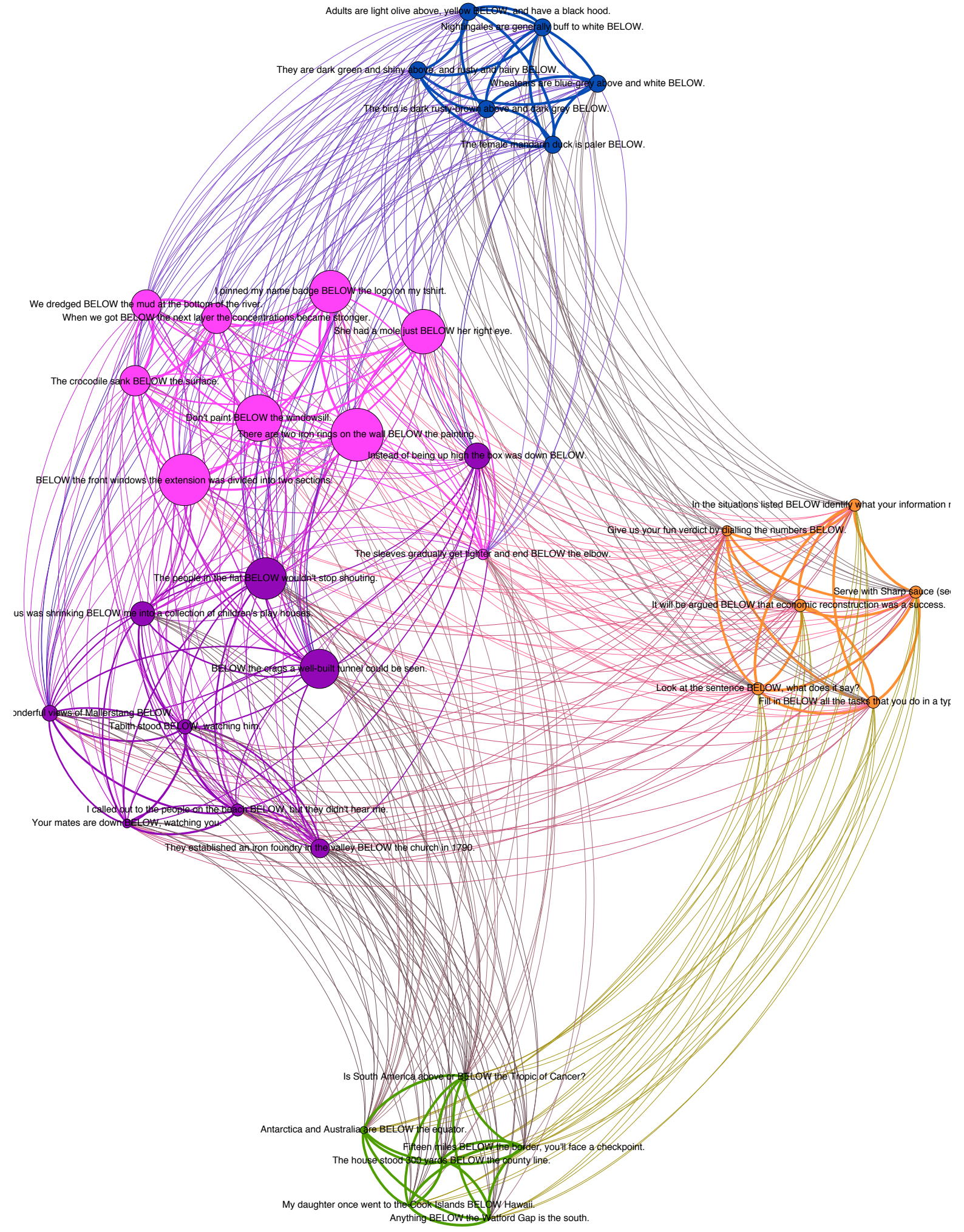
Below Non-spatial T1



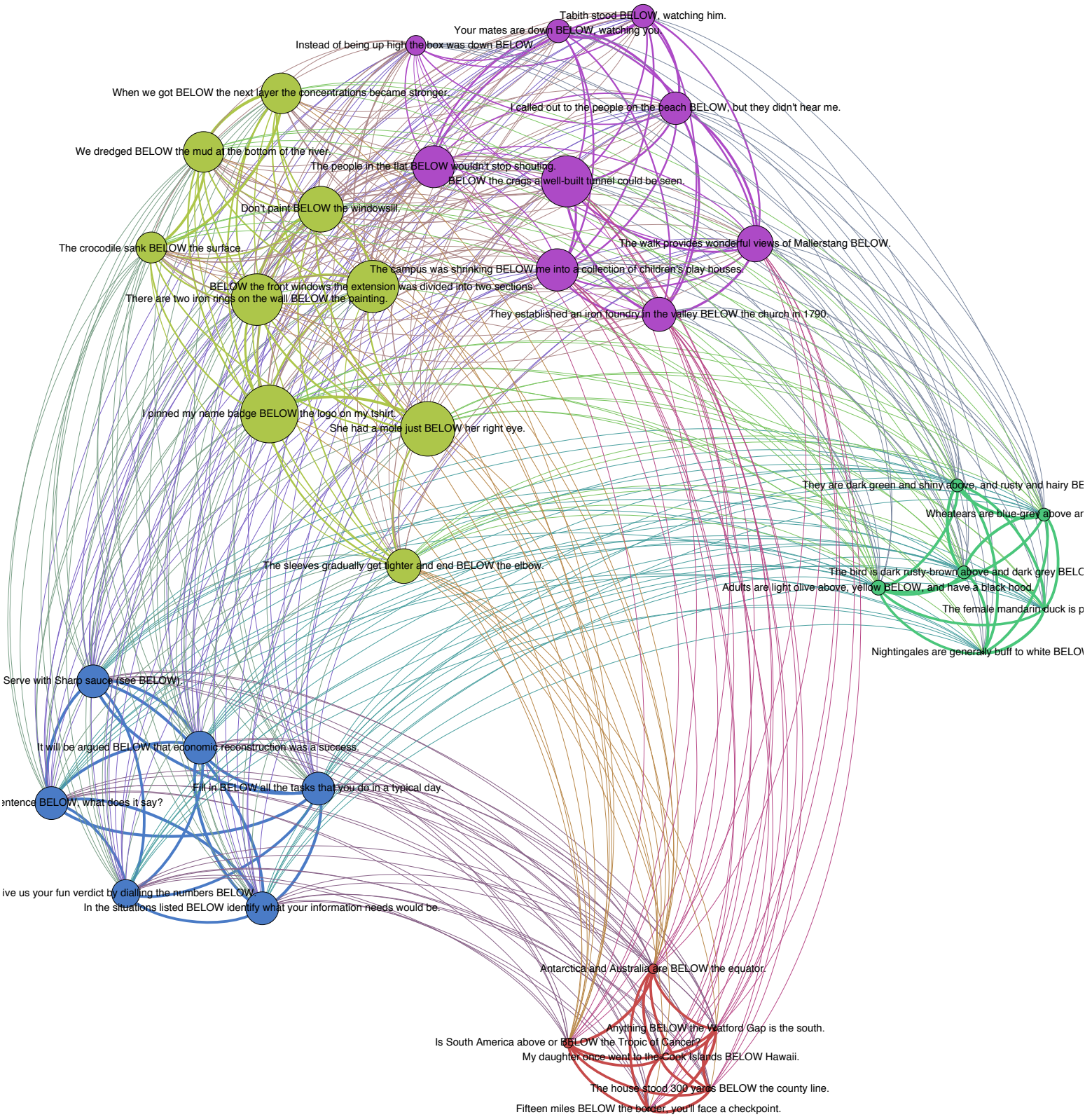
Below Non-spatial T2



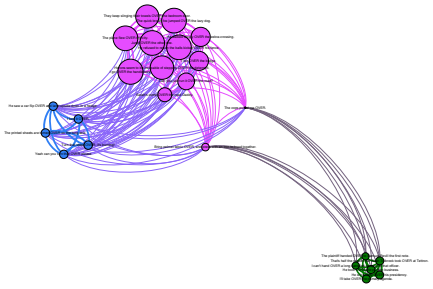
Below Spatial T1



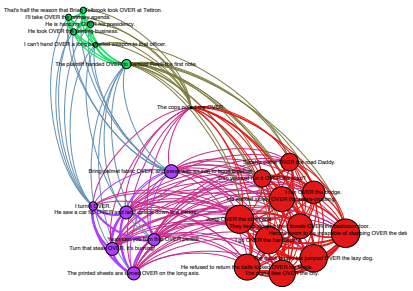
Below Spatial T2



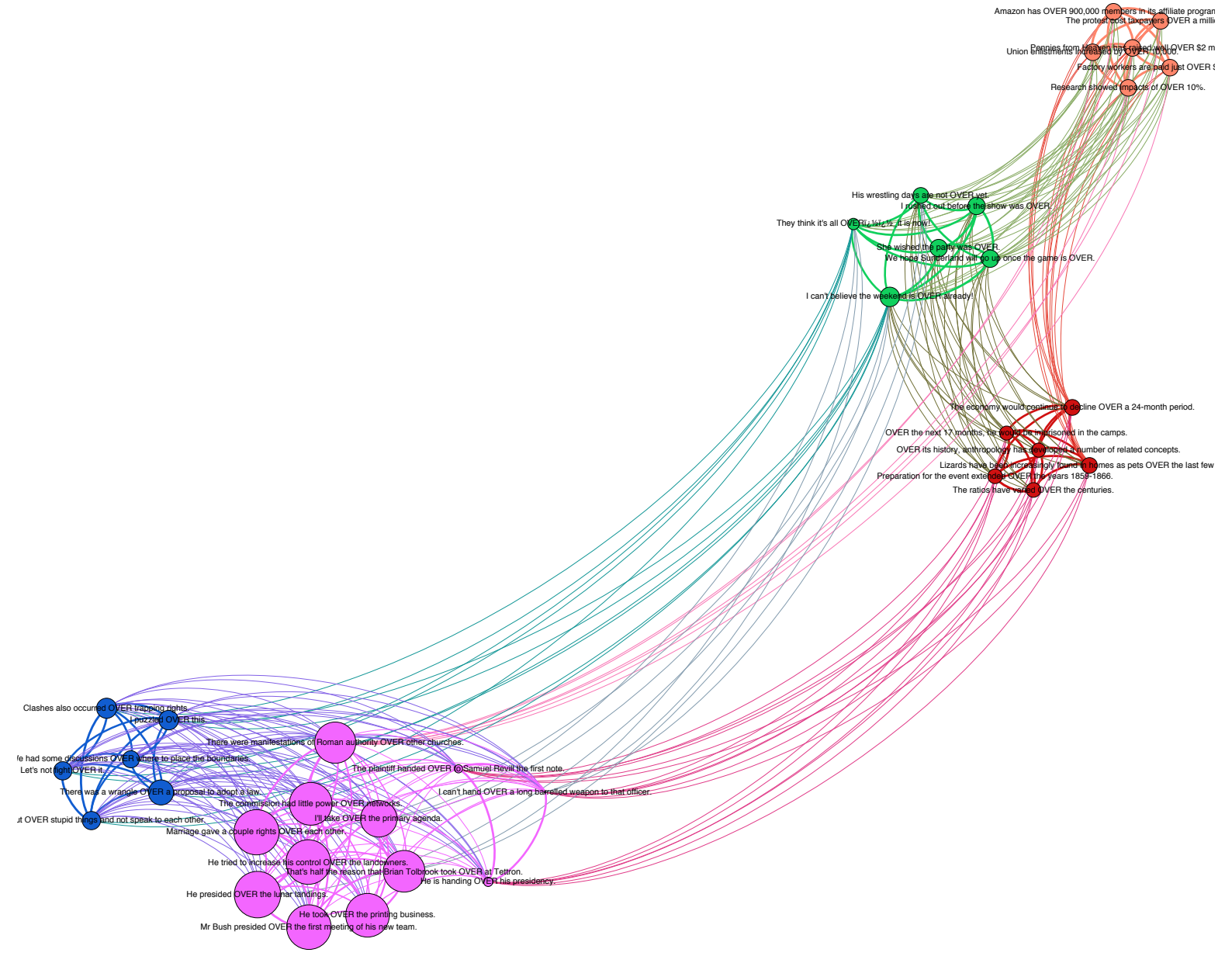
Over Mixed T1



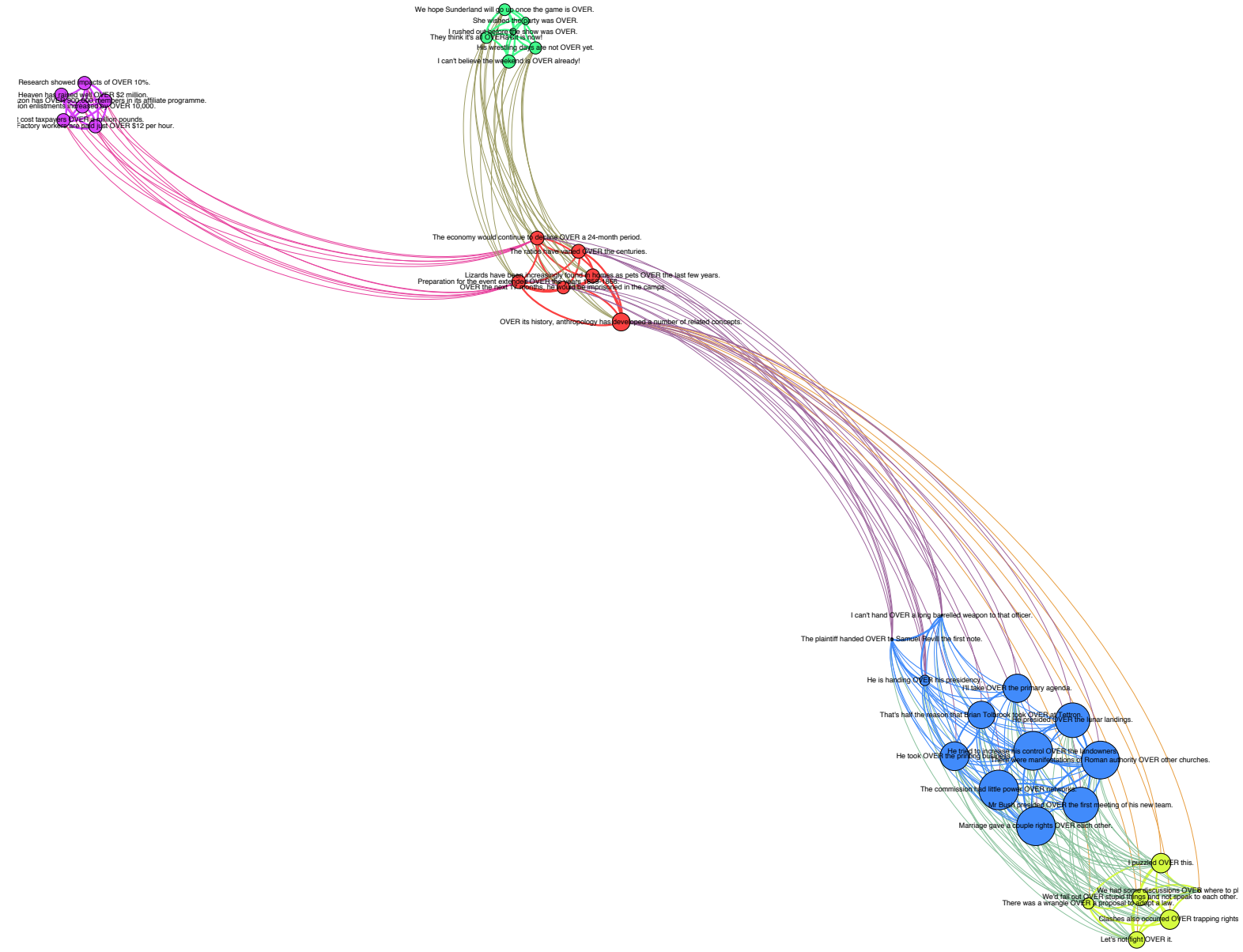
Over Mixed T2



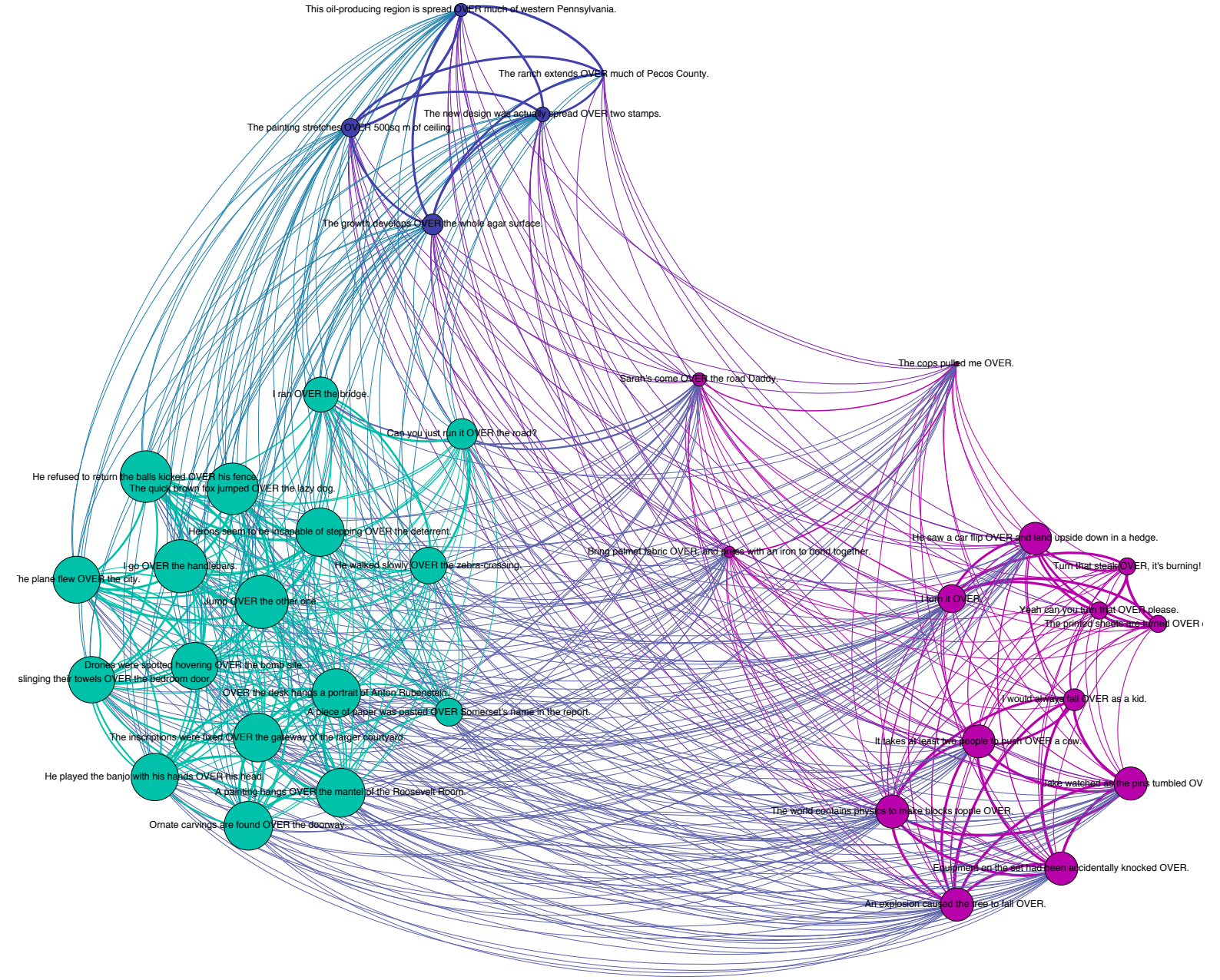
Over Non-spatial T1



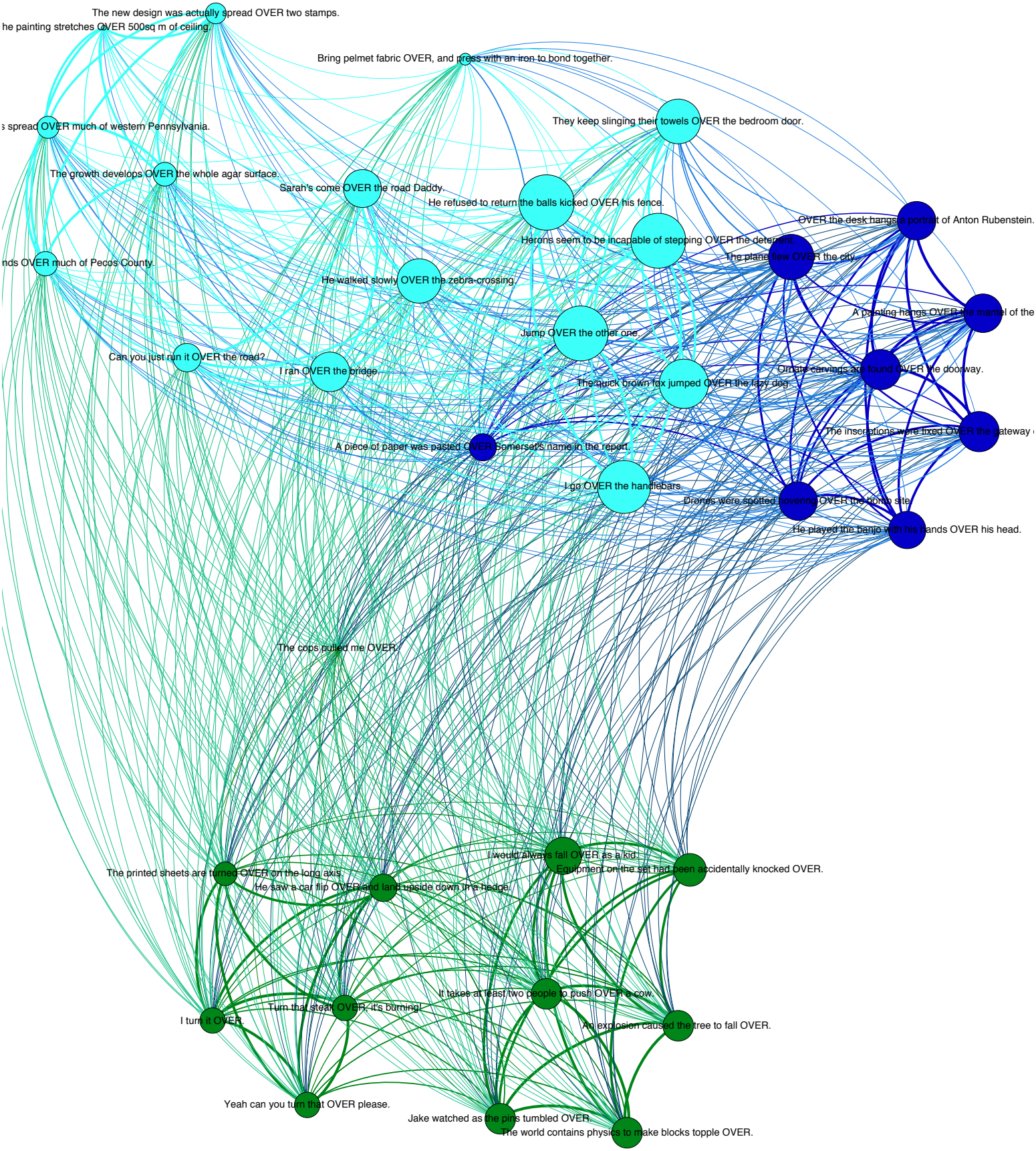
Over Non-spatial T2



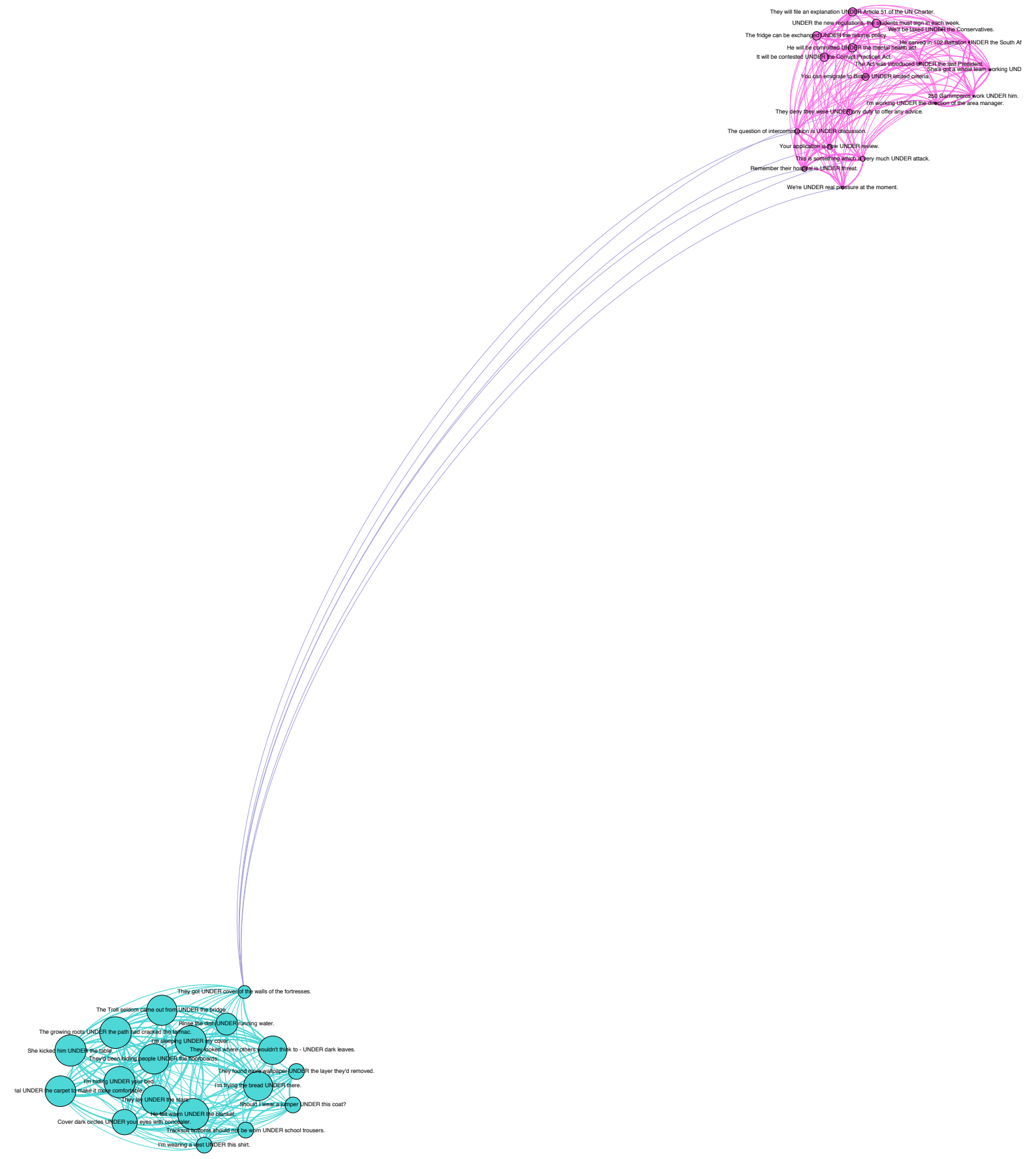
Over Spatial T1



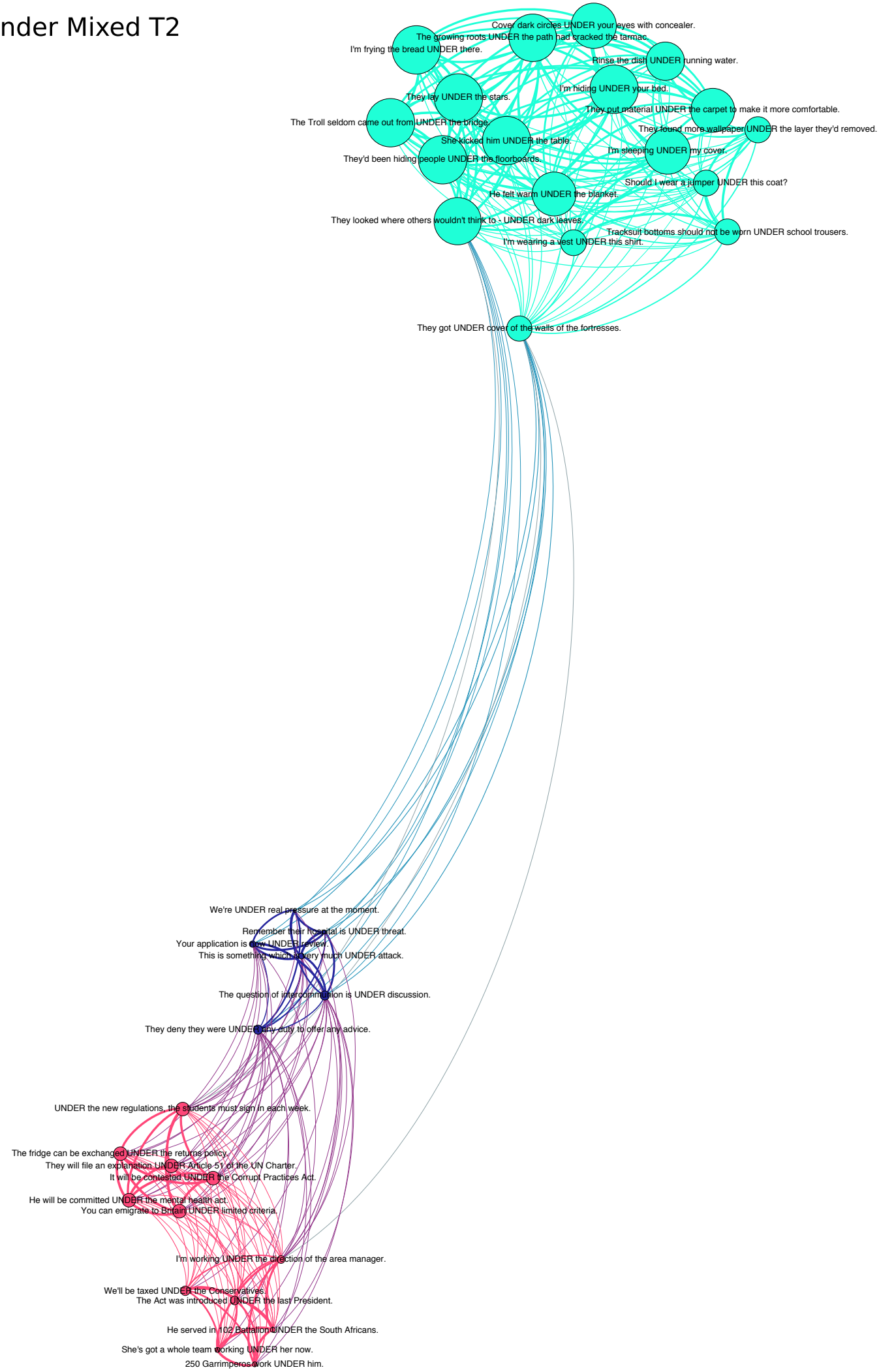
Over Spatial T2



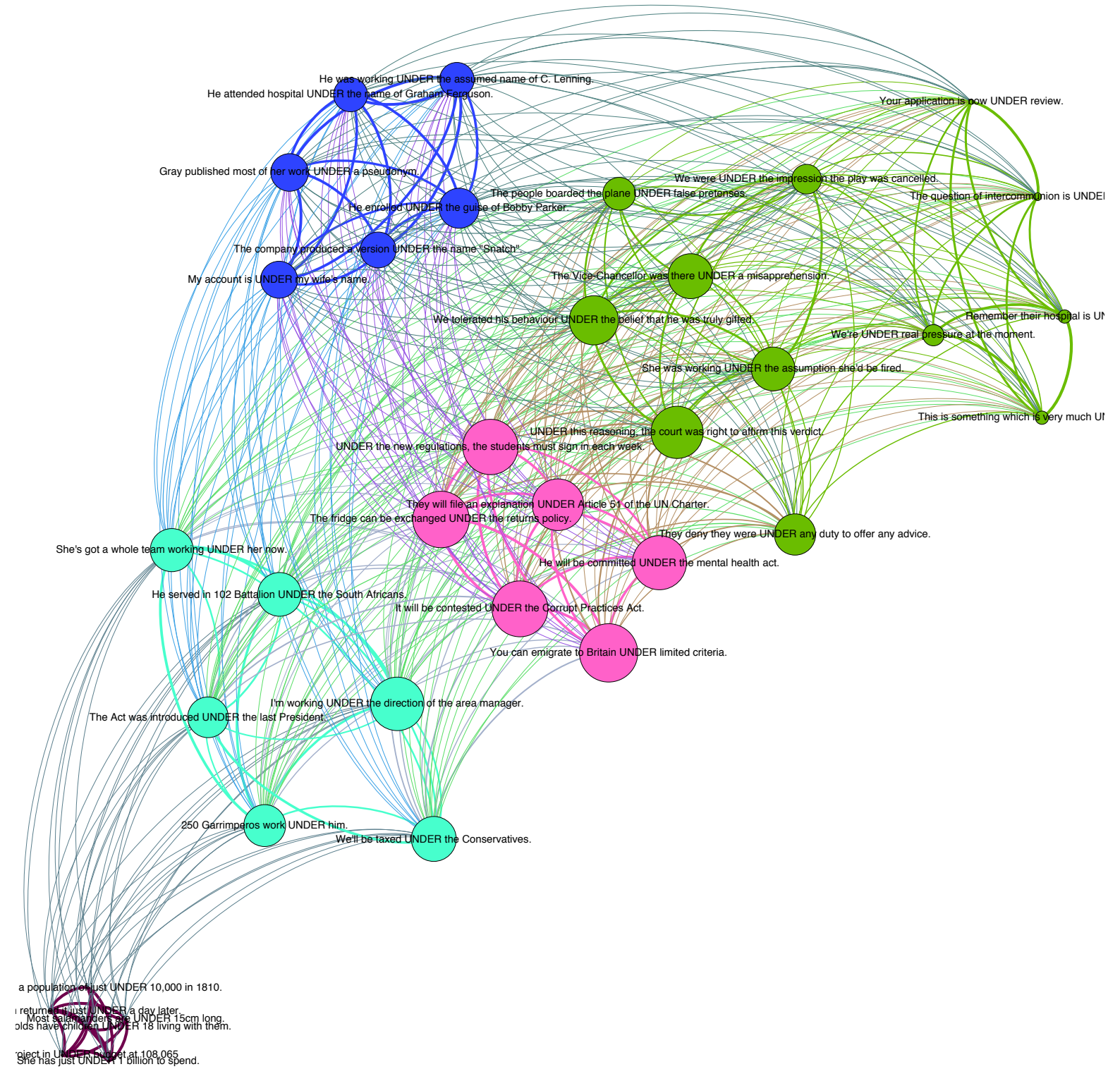
Under Mixed T1



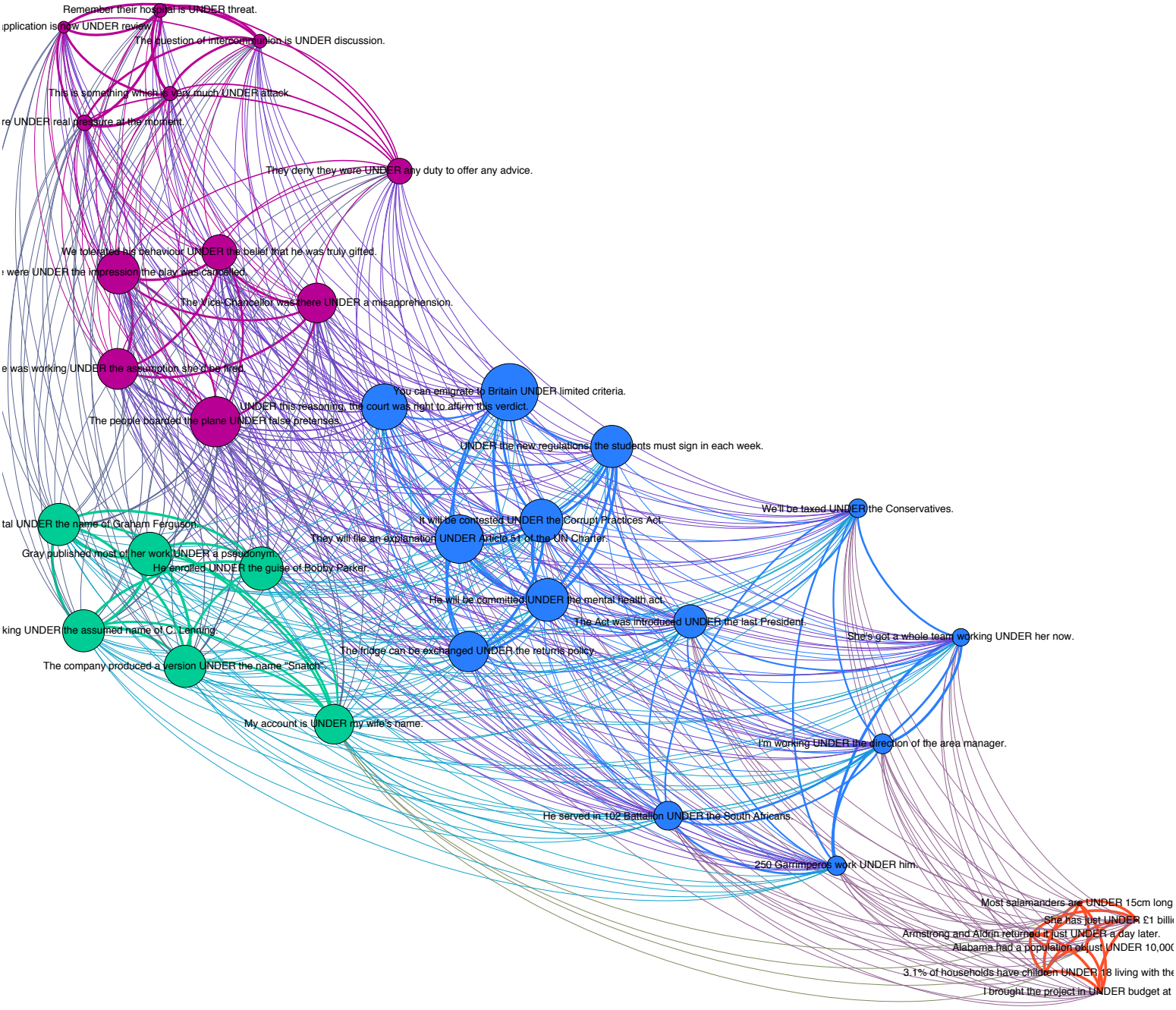
Under Mixed T2



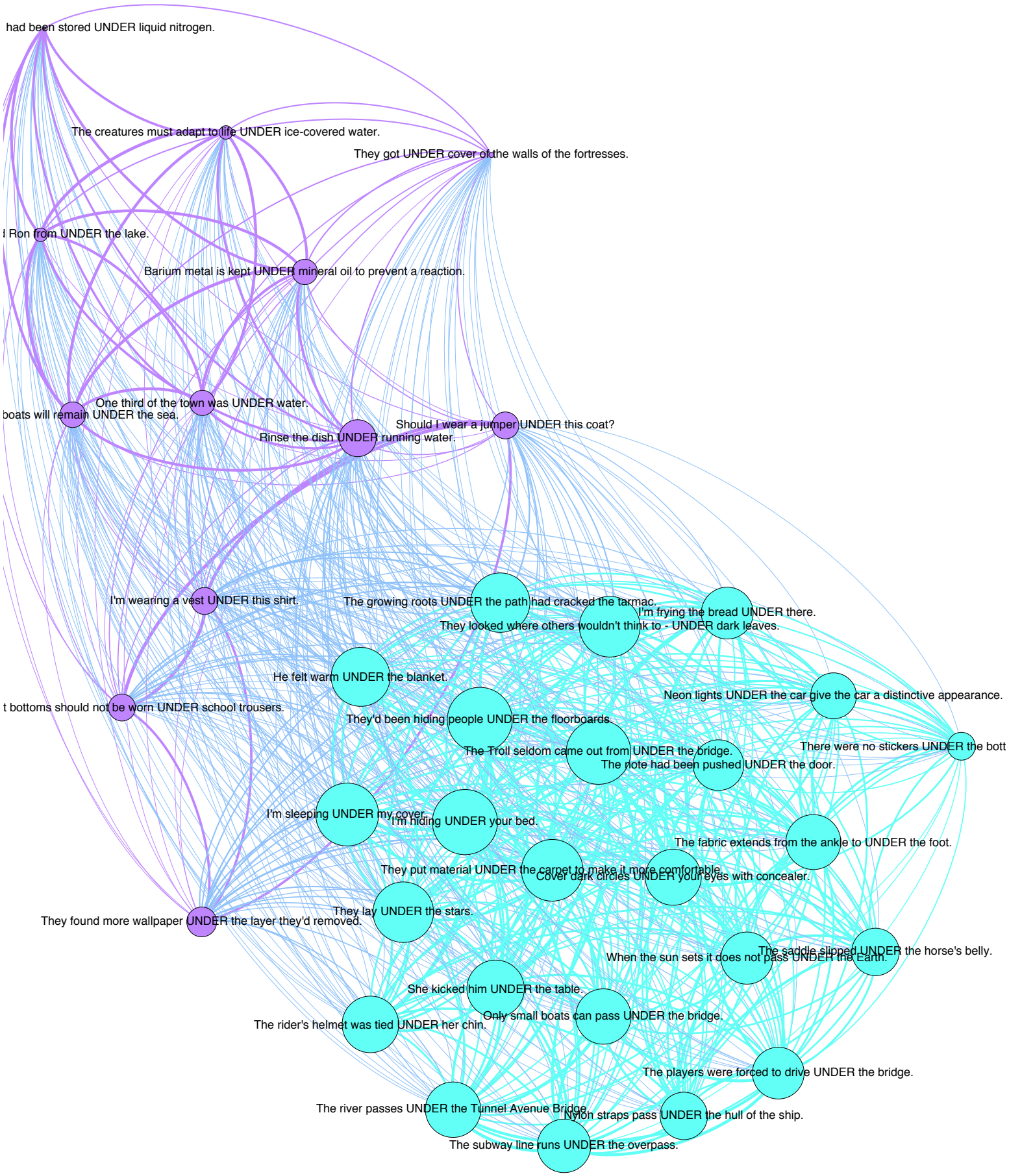
Under Non-spatial T1



Under Non-spatial T2



Under Spatial T1



Under Spatial T2